O presente livro contém trabalhos de pesquisadores dedicados a habilitar profissionais e pesquisadores a se sentirem mais seguros na elaboração dos seus próprios instrumentos de trabalho. Esperamos, ainda, que o livro encoraje os pesquisadores nacionais a enfrentarem tal empreendimento e de fato construírem os instrumentos de que tanto precisamos no país. Por esta razão, o livro foi planejado para servir de manual de suporte no desenvolvimento de instrumentos de medida e avaliação. Ainda não foi possível cobrirmos todos os tipos de instrumentos que o psicólogo pode fazer uso ou precisa para sua atuação como profissional ou pesquisador. Contudo, este livro constitui o início e esperamos que incentive o aparecimento de obras similares na área. Entretanto, o presente livro já cobre uma gama importante desta área, tratando em especial da teoria e da técnica de elaboração de testes psicológicos, de escalas psicométricas, de escalas psicofísicas para uso na avaliação de construtos psicológicos, do survey, de provas de avaliação do desempenho, do diferencial semântico, bem como da utilização da informática nesta área.

ISBN 85-900993-1-8

Luiz Pasquali

Instrumentos Psicológicos: Manual Prático de Elaboração

Instrumentos Psicológicos: Manual Prático de Elaboração

editado por
LUIZ PASQUALI

PAM / IBAPP

Esta obra é uma promoção das seguintes instituições:

LabPAM:

O Laboratório de Pesquisa em Avaliação e Medida, pertencente ao Departamento de Psicologia Social e do Trabalho do Instituto de Psicologia da UnB, foi fundado em 1987 para estudar a problemática da avaliação do comportamento humano, em suas mais variadas aplicações (educação, saúde, trabalho, organizações, trânsito, segurança pública), bem como para a criação de produtos e instrumentais que sejam úteis na prática profissional nessas mesmas áreas da atividade humana.

O LabPAM constitui a infra-estrutura para o grupo de pesquisa "Teoria e Metodologias de Medida e Avaliação". Este grupo é formado por uma equipe de pesquisadores ligados aos cursos Graduação, Mestrado e Doutorado coordenados pelos seguintes professores:

- Luiz Pasquali
  (Docteur, Université Catholique de Louvain)
- Bartholomeu T. Tróccoli
  (PhD. University of Wisconsin-Madison)
- Jacob A. Laros
  (PhD. Rijksuniversiteit Groningen)

São objetivos do Laboratório:

a) estudo da metodologia de pesquisa social;
b) construção, validação e adaptação de instrumentos de medida;
c) validação de programas educacionais e sociais.

IBAPP:

O Instituto Brasileiro de Avaliação e Pesquisa em Psicologia foi fundado em 1997 pelos pesquisadores brasileiros sob o patrocínio da Conselho Federal de Psicologia - CFP, com a finalidade de verificar e fundamentar as bases técnico-científicas dos instrumentos de avaliação psicológica no país.

O instituto criou uma revista especializada (Revista: *Avaliação Psicológica*) para a publicação de trabalhos sobre avaliação psicológica no Brasil.

INSTRUMENTOS PSICOLÓGICOS: MANUAL PRÁTICO DE ELABORAÇÃO

Editor
Luiz Pasquali
Diretoria do Laboratório de Pesquisa em Avaliação e Medida – LabPAM
Luiz Pasquali
Bartholomeu Tôrres Tróccoli
Jacob A. Laros
Diretoria do Instituto Brasileiro de Avaliação e Pesquisa em Psicologia – IBAPP
Presidente:
Luiz Pasquali
Vice-Presidente
Solange Wechsler
Secretário
Marcelo Tavares
Tesoureiro
Balsem Pinelli Jr.
Supervisão
Luiz Pasquali
Revisão
Luiz Pasquali
Bartholomeu Tôrres Tróccoli
Robson Medeiros de Araújo
Capa, Projeto Gráfico e Editoração Eletrônica
Print Laser Assessoria Editorial Ltda.
Impresso no Brasil por
Prática Gráfica e Editora Ltda.

INSTRUMENTOS PSICOLÓGICOS: MANUAL PRÁTICO DE ELABORAÇÃO

Luiz Pasquali
Organizador

Laboratório de Pesquisa em Avaliação e Medida – LabPAM
(*Instituto de Psicologia da Universidade de Brasília*)
Instituto Brasileiro de Avaliação e Pesquisa em Psicologia - IBAPP
Brasília - 1999

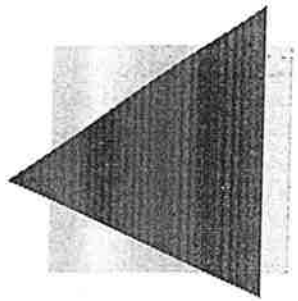
Laboratório de Pesquisa em Avaliação e Medida – LabPAM
Caixa Postal: 4464 – CEP: 70919-970 – Brasília - DF

Ficha Catalográfica:

I59 Instrumentos psicológicos: manual prático de elaboração / Luiz Pasquali (organizador). — Brasília : LabPAM; IBAPP, 1999.
306 p. : IL.
IBSN: 85-900993-1-8

1. Testes psicológicos. 2. Pesquisa em psicologia. I. Pasquali, Luiz.

CDU – 159.9.018.7



ÍNDICE

Prefácio 7
Autores 9
Cap. 1 Histórico dos instrumentos psicológicos 13
Luiz Pasquali
Cap. 2 Taxonomia dos instrumentos psicológicos 27
Luiz Pasquali
Cap. 3 Testes referentes a construto: Teoria e modelo de construção 37
Luiz Pasquali
Cap. 4 Utilização de escalas de razão de variáveis clínicas e sociais 73
Fátima Aparecida Emm Faleiros Sousa, Ricardo Kamizaki e José Aparecido Da Silva
Cap. 5 Escalas psicométricas 105
Luiz Pasquali
Cap. 6 O diferencial semântico 127
Luiz Pasquali
Cap. 7 Testes referentes a conteúdo: Medidas educacionais 141
Luiz Pasquali e Amélia Regina Alves
Anexo: Outras Taxionomias Educacionais 181
Cap. 8 Testes centrados em critério (CRT) 189
Leandro S. Almeida
Cap. 9 Tests informatizados 209
Gerardo Prieto Adánez
Cap. 10 Como elaborar um questionário 231
Hartmut Günther

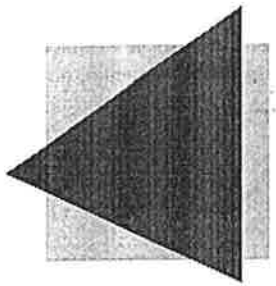
Cap. 11 Desenvolvimento das rotas de leitura fonológica e lexical em escolares, e de seu comprometimento em disléxicos 259
Fernando Cesar Capovilla, Elizeu Coutinho de Macedo, Marcelo Duduchi e Roberto Amilton Bernardes Sória

Anexo 1: O algoritmo para a determinação de *endpoints* em tempo real empregado no *software* CronoFonos 2.0 287

Anexo 2: Algumas perspectivas de aplicação de CronoFonos 2.0 para avaliação de desenvolvimento de leitura bem como de análise de padrões de déficits de leitura em dislexias 291

Índice por Autor 295

Índice por Assunto 303



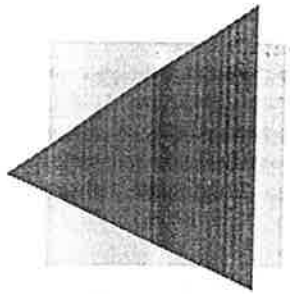
Prefácio

Cada vez mais, instrumentos de avaliação psicológica e de desempenho vêm sendo empregados no Brasil nas mais variadas áreas de atuação profissional, inclusive com crescente imposição legal, nomeadamente na esfera dos ditos psicotécnicos. Para o psicólogo pesquisador os instrumentos são essenciais para o avanço do conhecimento psicológico no país. Esta crescente demanda de instrumentos de avaliação e medida não está acompanhada por um desenvolvimento dos mesmos por parte dos pesquisadores brasileiros. Estes, na maioria dos casos, se contentam em produzir um trabalho sumário sobre instrumentos estrangeiros, sem maiores preocupações com a aferição da qualidade dos mesmos e aplicabilidade para o nosso contexto cultural. A causa principal disto deve-se à formação dos psicólogos, tipicamente carente de esforços maiores ou suficientes, para desenvolver, nos futuros profissionais, técnicas adequadas para uma avaliação criteriosa, muito menos para a elaboração de instrumentos de qualidade científica. O descrédito que os testes psicológicos, particularmente os ditos psicométricos, têm no Brasil se deve, ao que parece, principalmente a esta formação deficiente na área: muitos psicólogos no Brasil sequer se imaginam como profissionais aptos para criticar e construir seu material de trabalho, mesmo sabendo que os instrumentos constituem material de utilidade e necessidade cotidiana. Felizmente, estão surgindo no país grupos de pesquisadores especificamente dedicados a este problema, que, ainda que incipiente e pequeno em número, já começam, pelo menos, a incomodar a classe dos psicólogos do país com respeito ao problema da instrumentação psicológica, da sua qualidade, do seu uso e da sua melhoria.

O presente livro contém trabalhos de pesquisadores dedicados a habilitar profissionais e pesquisadores a se sentirem mais seguros na elaboração dos seus próprios instrumentos de trabalho. Esperamos, ainda, que o livro encoraje os pesquisadores nacionais a enfrentarem tal empreendimento e de fato construírem os instrumentos de que tanto precisamos no país. Por esta razão, o livro foi planejado para servir de manual de suporte no desenvolvimento de instrumentos de medida e avaliação. Ainda não foi possível cobrirmos todos os tipos de instrumentos que o psicólogo pode fazer uso ou precisa para sua atuação como profissional ou pesquisador. Contudo, este livro constitui o início e esperamos que incentive o aparecimento de obras similares na área. Entretanto, o presente livro já cobre uma gama importante

desta área, tratando em especial da teoria e da técnica de elaboração de testes psicológicos, de escalas psicométricas, de escalas psicofísicas para uso na avaliação de construtos psicológicos, do survey, de provas de avaliação do desempenho, do diferencial semântico, bem como da utilização da informática nesta área.

Luiz Pasquali



Os Autores

Amélia Regina Alves

Nascida no Rio de Janeiro, estudou psicologia na Universidade de Brasília. Ainda na graduação definiu contornos do tema de sua tese de mestrado sempre motivada por jovens professores do Instituto de Psicologia. Ingressou na Telecomunicações Brasileiras S/A, logo após a formatura ingressou como psicóloga no Centro Nacional de Treinamento da Telebrás, onde trabalhou por 19 anos, teve a oportunidade de supervisionar a área de tecnologia instrucional. Nos últimos anos, como coordenadora do Núcleo de Avaliação desenvolveu e implantou, em conjunto com a Universidade de Brasília, o Sistema de Avaliação do Treinamento. Dessa parceria surgiu uma importante linha de pesquisa, que resultou na publicação de alguns trabalhos relacionados a Resultados do Treinamento. Atualmente trabalha na Agência Nacional de Telecomunicações ANATEL, na área de Estudos de Mercado.

Elizeu Coutinho de Macedo

Nasceu no Rio de Janeiro (RJ) em 1968. Graduado em Psicologia pelo Instituto de Psicologia da USP (IPUSP) em 1991. Mestre em Psicologia Experimental pelo IPUSP em 1994 e Doutorando em Psicologia Experimental pelo mesmo instituto. Atualmente é professor da Universidade de Guarulhos (UNG). Publicou cerca de 40 trabalhos nas áreas de: avaliação neuropsicológica, reabilitação cognitiva, e desenvolvimento de instrumentos computadorizados de Comunicação Alternativa. Trabalhou no desenvolvimento de mais de uma centena de programas de computador no Laboratório de Neuropsicolingüística Experimental do IPUSP sob coordenação do Prof. Capovilla. É co-autor dos livros: *Manual ilustrado de sinais e sistemas de comunicação em rede para surdos* e *Tecnologia em (Re)Habilitação Cognitiva: Uma perspectiva multidisciplinar.* Membro da *International Society for Augmentative and Alternative Communication* (ISAAC).

Fátima A.E. Faleiros Sousa

Nascida em Ribeirão Preto (SP) em 1956. Tem formação em Enfermagem. Obteve o doutorado em Enfermagem pela Escola de Enfermagem de Ribeirão Preto

da USP. Atualmente é Professor Associado nessa mesma instituição, na qual é docente desde 1986. Desde a defesa da tese de doutorado, as pesquisas da autora são referentes à mensuração de variáveis subjetivas (clínicas e sociais) através de métodos rigorosos oriundos da Psicofísica Sensorial. Para desenvolver projetos experimentais dessa natureza, recebeu auxílio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) para a construção do Laboratório de Psicofísica Clínica. Publicou vários trabalhos na área. Atualmente tem-se dedicado em particular à Mensuração e Avaliação da Dor (Psicofísica da Dor).

Fernando C. Capovilla

Nasceu em Campinas (SP) em 1960. É Ph.D. em Psicologia Experimental Humana pela Temple University, Philadelphia, PA, USA. No Instituto de Psicologia da Universidade de São Paulo, fundou e coordena o Laboratório de Neuropsicolingüística Cognitiva Experimental, o Laboratório de Tecnologia para Avaliação e Reabilitação Cognitivas, e o Centro de Atendimento em Deficiências Sensoriais e Distúrbios Motores e Cognitivos no Cérebro-Lesado. Nesses laboratórios e centro de atendimento, coordenou a pesquisa e desenvolvimento de mais de uma centena de sistemas computadorizados de diagnóstico, tratamento, reabilitação e comunicação de pacientes com deficiências sensoriais e distúrbios motores e de processamento cognitivo, cuja aplicação na clínica experimental resultou em várias centenas de pacientes atendidos e em mais de 80 artigos científicos e 25 capítulos de livro publicados. É co-autor dos livros: *Manual ilustrado de sinais e sistema de comunicação em rede para surdos*, *Reabilitação Cognitiva: uma perspectiva multidisciplinar*, e *Dicionário da língua brasileira de sinais: ilustração, descrição e escrita visual direta de 3500 sinais*, além de dois livros sobre avaliação e tratamento de distúrbios de leitura, no prelo. Ocupa o cargo de *President-Elect* (1996-2000) do *Brazilian Chapter* da *International Society for Alternative Communication*, uma associação internacional dedicada à pesquisa e desenvolvimento de recursos para portadores de distúrbios de comunicação oral, escrita e de sinais.

Gerardo Prieto Adanez

Nasceu en Zamora (Espanha). Cursou Psicologia na Universidad de Salamanca, obtendo o doutorado na mesma Universidade em 1977. Atualmente é Catedrático da Universidad de Salamanca onde tem lecionado Estatística aplicada à Psicologia e Psicometria desde 1974. É autor de mais de 80 trabalhos, publicados em revistas hispanas e internacionais, relacionados com a construção de testes e a avaliação das aptidões. Nos últimos anos, coordenou vários projetos de investigação financiados por diversas instituições hispanas, tanto públicas como privadas. Deve-se destacar sua dedicação ao campo dos testes informatizados, em que participou desenvolvendo baterias informatizadas para a seleção de pilotos, avaliação de aptidões, espacial, etc. Recentemente coeditou o livro *Tests Informatizados: fundamentos y aplicaciones*, publicado pela editora espanhola Pirámide.

Hartmut Günther

Nascido na Austria, estudou na Universität Hamburg e na Universität Marburg. Graduou-se em Psicologia pelo Albion College, em Michigan, USA. Tendo realizado o mestrado em Psicologia Experimental na Western Michigan University e o doutorado em Psicologia Social na University of California at Davis, é atualmente pesquisador e professor titular da Universidade de Brasília, onde fundou e coordena o Laboratório de Psicologia Ambiental, sendo o coordenador do Programa de Pós-Graduação do Instituto de Psicologia. Tem trabalhos publicados no Brasil e no exterior, em especial sobre metodologia e a psicologia ambiental.

José Aparecido Da Silva

Natural de Jaboticabal (SP) em 1952. Mestre e Doutor pelo IPUSP e Professor Visitante na University of California, Santa Barbara, USA. Atualmente é Professor Titular do departamento de Psicologia e Educação da FFCLRP-USP e Diretor dessa Instituição. É coordenador científico da área de Psicologia do CNPq. Tem diversos artigos publicados em revistas nacionais e internacionais indexadas. Muitos de seus trabalhos têm sido freqüentemente citados na literatura internacional, sendo um dos pesquisadores em Psicologia mais freqüentemente citados pelos seus pares, tanto em periódicos quanto em livros e capítulos de livros. Seu principal interesse em pesquisa reside na análise dos processos perceptuais e cognitivos subjacentes ao comportamento espacial. Os problemas básicos que têm sido investigados são: percepção espacial, controle visual da ação, psicofísica social e clínica e teoria geral da mensuração (escalas e testes). Um grande número de artigos tem sido publicado em revistas muito prestigiosas na área, como: Journal of Experimental Psychology: *Human Perception and Performance; Perception & Psychophysics, Current Directions in Psychological Science*. É o único pesquisador brasileiro membro da Psychonomic Society e pesquisador com dados bibliográficos incluídos no *Who's is Who in the World*, 1988 e na edição especial do ano 2000.

Leandro S. Almeida

Natural do Porto, Portugal. Doutorado em Psicologia, especialidade de Psicologia da Educação, pela Universidade do Porto (1987). Atualmente é Professor Titular do Instituto de Educação e Psicologia da Universidade do Minho. Investigações nos domínios da inteligência, desenvolvimento cognitivo e aprendizagem escolar. Coordenou a padronização de algumas provas psicológicas no domínio da inteligência para a população portuguesa. Foi presidente da Associação dos Psicólogos Portugueses e coordenador da Divisão de Psicologia Escolar desta Associação. Coordena em Portugal a realização da Conferência Internacional sobre "Avaliação Psicológica: Formas e Contextos". São da sua autoria os livros *Teoria da Inteligência* e *O Raciocínio Diferencial dos Jovens: Avaliação, Desenvolvimento e Diferenciação*, além de vários artigos sobre testes e educação.

Luiz Pasquali

Nasceu em Gaurama (RS) em 1933. Tem formação em Pedagogia, Filosofia e Psicologia. Obteve o doutorado em Psicologia pela Université Catholique de Louvain em 1970. Lecionou em Michigan nos USA, na PUCRS e atualmente é Professor Titular no Instituto de Psicologia da UnB, onde leciona desde 1975. Desde sua tese de doutorado, o interesse e as pesquisas do autor se dirigem para

o problema da instrumentação e da medida em ciências psicossociais, em particular na Psicologia, criando, através da FINEP, um laboratório de pesquisa nesta área na UnB, bem como um laboratório de ensino de disciplinas via computador. Publicou cerca de 50 trabalhos na área da instrumentação psicológica, sendo autor e organizador do livro *Teoria e Métodos de Medida em Ciências do Comportamento* e autor do livro *Psicometria: Teoria e Aplicações*, tendo outros no prelo ou em fase de conclusão. Seu laboratório é membro da *International Test Commission* e é termo de referência na área da instrumentação psicológica no país e na Íbero-américa.

Marcelo Duduchi Feitosa

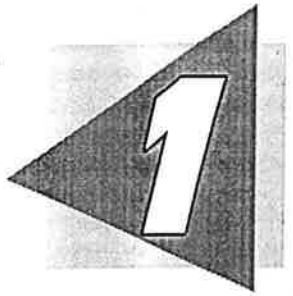
Nasceu no Rio de Janeiro (RJ) em 1967. Graduado em Tecnologia em Processamento de Dados pela Faculdade de Tecnologia de São Paulo (FATEC/SP) em 1989. Especialista em Automação Industrial pelo Centro de Desenvolvimento Tecnológico de São José dos Campos em 1991. Mestre em Neurociências e Comportamento pelo IPUSP em 1998, e doutorando em Psicologia Experimental pelo mesmo instituto. Atualmente é Professor da FATEC-SP e diretor dos cursos de Tecnologia em Processamento de Dados e de Ciência da Computação da Universidade da Cidade de São Paulo (UNICID). Publicou cerca de 30 trabalhos sobre desenvolvimento de sistemas computadorizados para uso em pesquisa e clínica neuropsicológica. Trabalhou no desenvolvimento de mais de uma centena de programas de computador no Laboratório de Neuropsicolingüística Experimental do IPUSP sob coordenação do Prof. Capovilla.

Ricardo Kamizaki

Nascido em Londrina (PR) em 1957. Tem formação em Psicologia e Mestrado pela Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo. Atualmente é aluno de Doutorado em Psicobiologia nesta mesma instituição e bolsista pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). Tem trabalhos publicados na área da Psicofísica Clínica e Social.

Roberto Amilton Bernardes Soria

Nasceu no Rio de Janeiro (RJ) em 1972. Graduado em Engenharia Eletrônica pela Escola Politécnica da USP em 1995. Mestrado em Sistemas Eletrônicos na POLI-USP. Publicou cerca de 12 trabalhos em redes neurais artificiais e processamento de voz.



Histórico dos Instrumentos Psicológicos*

Luiz Pasquali

1. Introdução

Os instrumentos psicológicos ou os testes, como os conhecemos hoje em dia, representam a expressão cientificamente sofisticada de um procedimento sistemático de qualquer organismo, biológico ou social, a saber, o de avaliar as situações para tomar decisões que garantam a sobrevivência do próprio organismo, bem como seu autodesenvolvimento. Aparece tal fenômeno desde o chamado crivo biológico, que filtra os estímulos para que possam ser adequadamente elaborados pelo organismo, até as formas mais sofisticadas dos testes psicológicos. Assim, embora o termo avaliação ("assessment") tenha uma história muito recente, seu uso, formal ou informal, data da origem dos organismos vivos. Ao que parece (Sundberg, 1977), a expressão avaliação ("assessment") apareceu como termo psicológico no livro "Assessment of Men" do *U.S. Office of Strategic Services*, em 1948, utilizado para expressar o conjunto de processos que as pessoas usam para formar impressões e imagens, tomar decisões e verificar hipóteses sobre as características das outras pessoas no confronto delas com seu meio ambiente. Desta forma, avaliar parece ser uma fatalidade do ser humano com relação ao seu meio ambiente, incluindo ali o meio físico bem como o social. Ele se constitui num processo informal e formal. Informal, na medida em que todo indivíduo avalia seu meio ambiente, os outros indivíduos, fazendo deles representações para em cima delas tomar decisões de como agir, no sentido de manter a própria sobrevivência e seu autodesenvolvimento. No que se refere ao aspecto formal da avaliação, a história humana está cheia de códigos de conduta através dos quais as pessoas e a sociedade julgavam e julgam o comportamento dos seus semelhantes. Estes julgamentos são feitos em termos de valores pessoais ou de valores impostos pela sociedade, sempre com a suposição de que respeito e conformidade com tais valores é algo positivo, sendo sua violação algo errado e condenável e, portanto, que deve ser reparado, tratado (via morte, cadeia, castigo, terapia). Os paradigmas da avaliação formal e do conseqüente controle do

* parte deste capítulo foi publicada no livro Pasquali, L. (1998). Psicometria: Teoria e aplicações. Brasília: Editora UnB.

comportamento dos indivíduos na sociedade variaram muito durante a história da humanidade, dependendo das crenças, filosofia e códigos de conduta de diferentes contextos culturais, desde o sistema familiar do período neolítico (12.000 antes de Cristo), do sistema de adivinhos das culturas egípcia e suméria (10.000 a.C.), até os testes psicológicos atuais (Barclay, 1991).

Entretanto, são muito escassos os relatos sobre o uso de técnicas de avaliação sistemática do comportamento humano nestas épocas longínquas. Dubois (1970) fala do uso de testes para seleção de funcionários civis da China lá por 3.000 a.C. Sabemos que o romano Galeno (116-200 A.D.) elaborou a teoria dos temperamentos que serviu muito tempo como paradigma de classificação da personalidade. Baseado e complementando a classificação médica de Hipócrates (5° século a.C.) dos quatro elementos da teoria dos fluidos (fogo, ar, água, terra), Galeno desenvolveu a teoria dos quatro temperamentos: melancólico (sujeito triste, reservado e sério), colérico (sujeito agressivo, excitável e impaciente), fleumático (sujeito passivo, pacífico e calmo) e sangüíneo (sujeito sociável, extrovertido e alegre). De fato, a origem mais sistemática dos testes psicológicos atuais pode ser traçada até o interesse dos estudiosos com a questão do número, desde a Idade Média, que iria desembocar nos psicólogos do fim do século passado na Alemanha, Inglaterra, França e Estados Unidos, onde se encontram as origens formais da psicometria. O estudo do número permitiu esta evolução por ser ele um símbolo que representa quantidade / extensão e que servia perfeitamente para basear técnicas de mensuração, que veio a ser a base da futura tecnologia de avaliação. Com isto fica igualmente acertado que a origem da avaliação formal e sistemática se deve aos psicólogos do fim do século passado.

De fato, estamos nos limiares do século XX. Em Psicologia vigoravam várias tendências epistemológicas, um tanto isoladas umas das outras, procurando superar o status pré-científico no estudo do psíquico (Boring, 1957).

A tradicional diatribe de origem cartesiana, alma vs. corpo, subsidiava estas tendências. Assim, temos, de um lado, a psicologia alemã da introspeção interessada na experiência subjetiva e, do outro, o empirismo inglês e norte-americano interessado no comportamento, bem como a escola (psicofísica) de Leipzig estudando os processos sensoriais. Estas duas grandes orientações se caracterizavam também pelo uso de procedimentos mais descritivos, no caso da psicologia introspectiva, e a procura de procedimentos mais quantitativistas por parte dos empiricistas. Portanto, não causa surpresa que as origens da Psicometria se encontrem no enfoque empiricista das psicologias da época. Desta sua origem, a Psicometria, tanto clássica quanto moderna (Teoria de Resposta ao Item - TRI) retém algumas caracterizações que permitem controvérsias. Entre elas, duas parecem particularmente fortes e, quiçá, preocupantes. Por um lado, a Psicometria, pelo menos na sua prática, é ainda guiada pela concepção positivista baconiana do empirismo (Bacon, 1620, 1984), isto é, que a ciência do universal se faz através do conhecimento do singular (indução), enfoque demonstrado como logicamente inviável, tanto pelo empiricista Hume (1739-1740) quanto por Popper (1972). Esta concepção creio ser responsável pelo descuido inaceitável da Psicometria com relação à teoria psicológica que deveria ser a preocupação preliminar e primordial na medida do psicológico. Preocupação esta que, felizmente, a Psicologia Cognitiva moderna procura ressuscitar. Por outro lado, predomina em Psicometria a concepção estatística sobre a psicológica. Os precursores e os que desenvolveram a Psicometria eram estatísticos de formação, tanto que ainda se define Psicometria como um ramo da Estatística, quando na verdade ela deve ser concebida como um ramo da Psicologia que interfaceia com a Estatística. Thurstone (1937) parecia preocupado com este problema, quando definiu como objeto de estudo para a sociedade psicométrica que acabara de fundar a "Psicologia Matemática", esta concebida como ramo da Psicometria dedicada à pesquisa dos modelos matemáticos dos processos psicológicos, mas sempre a serviço destes. De fato, a recente estatística (começou com Quetelet, 1796-1874) foi decisiva na elaboração da Psicometria e era entendido que através dela se podia avaliar adequadamente as diferenças individuais, objeto específico de qualquer avaliação. Temos neste contexto nomes e expoentes importantes tanto na Estatística quanto na Psicometria, como os de Sir Francis Galton (1822-1911), do seu discípulo Karl Pearson (1857-1936), Spearman (dos anos 1900) e Thurstone (dos anos 30). Esta plêiade de pesquisadores enviesou a Psicometria para o lado da Estatística, senão este que deve ser devidamente reparado no desenvolvimento daquela disciplina psicológica.

2. Origem da Psicometria

2.1 Apanhado histórico

A Psicometria (mais especificamente, os testes psicológicos) poderia ter tido origem em duas situações bastante distintas: 1) a psicologia de orientação empiricista ou 2) a psicologia mais mentalista de Binet na França. De fato, as duas tendências entraram em cena na mesma época para resolver os mesmos problemas, a saber, avaliar objetivamente as aptidões humanas. Apenas, Binet e Simon (1905) utilizando processos mentais e Galton (1883), Spearman (1904b) e outros empiricistas fazendo uso de processos comportamentais, mais especificamente, sensoriais. Embora o teste de inteligência de Binet tenha tido grande sucesso na Psicologia, não foi sua orientação que deu de fato a origem e o desenvolvimento à Psicometria porque lhe faltava o enfoque primordial da quantificação, que era o específico da orientação da psicologia empiricista. Da psicologia introspecionista da época realmente não se poderia esperar a origem da Psicometria, dada sua orientação puramente descritiva dos processos psicológicos. Conta Joncich (1968) que Thorndike (1904) ao enviar seu trabalho de medida em Psicologia a William James (que era da orientação descritiva) incluiu uma nota dizendo que o manuscrito era para seus alunos e que não aconselhava ao próprio James sua leitura!

Assim, a origem da Psicometria deve ser procurada nos trabalhos do estatístico Spearman (1904a, 1904b, 1907, 1913), que, no que se refere à Psicologia, seguiu os procedimentos fisicalistas de Galton (1883). Também não se deve estranhar que a Psicometria surgisse no campo das aptidões humanas (mentais, físicas, psicofísicas), pois, além de ser a temática psicológica da época, elas se coadunavam melhor a um estudo quantitativo, pois pode-se ali contabilizar o comportamento em termos de acertos e erros.

Aliás, para melhor entender a origem da Psicometria, pode-se seguir duas orientações, de início bastante independentes, que mais tarde se unificariam no que podemos chamar da Psicometria Clássica, a saber, a preocupação mais prática da

o problema da instrumentação e da medida em ciências psicossociais, em particular na Psicologia, criando, através da FINEP, um laboratório de pesquisa nesta área na UnB, bem como um laboratório de ensino de disciplinas via computador. Publicou cerca de 50 trabalhos na área da instrumentação psicológica, sendo autor e organizador do livro *Teoria e Métodos de Medida em Ciências do Comportamento* e autor do livro *Psicometria: Teoria e Aplicações*, tendo outros no prelo ou em fase de conclusão. Seu laboratório é membro da *International Test Commission* e é termo de referência na área da instrumentação psicológica no país e na Íbero-américa.

Marcelo Duduchi Feitosa

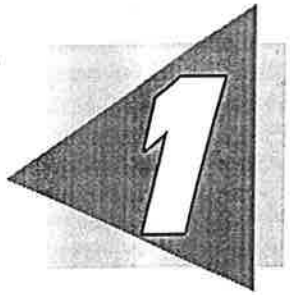
Nasceu no Rio de Janeiro (RJ) em 1967. Graduado em Tecnologia em Processamento de Dados pela Faculdade de Tecnologia de São Paulo (FATEC/SP) em 1989. Especialista em Automação Industrial pelo Centro de Desenvolvimento Tecnológico de São José dos Campos em 1991. Mestre em Neurociências e Comportamento pelo IPUSP em 1998, e doutorando em Psicologia Experimental pelo mesmo instituto. Atualmente é Professor da FATEC-SP e diretor dos cursos de Tecnologia em Processamento de Dados e de Ciência da Computação da Universidade da Cidade de São Paulo (UNICID). Publicou cerca de 30 trabalhos sobre desenvolvimento de sistemas computadorizados para uso em pesquisa e clínica neuropsicológica. Trabalhou no desenvolvimento de mais de uma centena de programas de computador no Laboratório de Neuropsicolingüística Experimental do IPUSP sob coordenação do Prof. Capovilla.

Ricardo Kamizaki

Nascido em Londrina (PR) em 1957. Tem formação em Psicologia e Mestrado pela Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da Universidade de São Paulo. Atualmente é aluno de Doutorado em Psicobiologia nesta mesma instituição e bolsista pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). Tem trabalhos publicados na área da Psicofísica Clínica e Social.

Roberto Amilton Bernardes Soria

Nasceu no Rio de Janeiro (RJ) em 1972. Graduado em Engenharia Eletrônica pela Escola Politécnica da USP em 1995. Mestrado em Sistemas Eletrônicos na POLI-USP. Publicou cerca de 12 trabalhos em redes neurais artificiais e processamento de voz.



Histórico dos Instrumentos Psicológicos*

Luiz Pasquali

1. Introdução

Os instrumentos psicológicos ou os testes, como os conhecemos hoje em dia, representam a expressão cientificamente sofisticada de um procedimento sistemático de qualquer organismo, biológico ou social, a saber, o de avaliar as situações para tomar decisões que garantam a sobrevivência do próprio organismo, bem como seu autodesenvolvimento. Aparece tal fenômeno desde o chamado crivo biológico, que filtra os estímulos para que possam ser adequadamente elaborados pelo organismo, até as formas mais sofisticadas dos testes psicológicos. Assim, embora o termo avaliação ("assessment") tenha uma história muito recente, seu uso, formal ou informal, data da origem dos organismos vivos. Ao que parece (Sundberg, 1977), a expressão avaliação ("assessment") apareceu como termo psicológico no livro "Assessment of Men" do *U.S. Office of Strategic Services*, em 1948, utilizado para expressar o conjunto de processos que as pessoas usam para formar impressões e imagens, tomar decisões e verificar hipóteses sobre as características das outras pessoas no confronto delas com seu meio ambiente. Desta forma, avaliar parece ser uma fatalidade do ser humano com relação ao seu meio ambiente, incluindo ali o meio físico bem como o social. Ele se constitui num processo informal e formal. Informal, na medida em que todo indivíduo avalia seu meio ambiente, os outros indivíduos, fazendo deles representações para em cima delas tomar decisões de como agir, no sentido de manter a própria sobrevivência e seu autodesenvolvimento. No que se refere ao aspecto formal da avaliação, a história humana está cheia de códigos de conduta através dos quais as pessoas e a sociedade julgavam e julgam o comportamento dos seus semelhantes. Estes julgamentos são feitos em termos de valores pessoais ou de valores impostos pela sociedade, sempre com a suposição de que respeito e conformidade com tais valores é algo positivo, sendo sua violação algo errado e condenável e, portanto, que deve ser reparado, tratado (via morte, cadeia, castigo, terapia). Os paradigmas da avaliação formal e do conseqüente controle do

* parte deste capítulo foi publicada no livro Pasquali, L. (1998). Psicometria: Teoria e aplicações. Brasília: Editora UnB.

comportamento dos indivíduos na sociedade variaram muito durante a história da humanidade, dependendo das crenças, filosofia e códigos de conduta de diferentes contextos culturais, desde o sistema familiar do período neolítico (12.000 antes de Cristo), do sistema de adivinhos das culturas egípcia e suméria (10.000 a.C.), até os testes psicológicos atuais (Barclay, 1991).

Entretanto, são muito escassos os relatos sobre o uso de técnicas de avaliação sistemática do comportamento humano nestas épocas longínquas. Dubois (1970) fala do uso de testes para seleção de funcionários civis da China lá por 3.000 a.C. Sabemos que o romano Galeno (116-200 A.D.) elaborou a teoria dos temperamentos que serviu muito tempo como paradigma de classificação da personalidade. Baseado e complementando a classificação médica de Hipócrates (5º século a.C.) dos quatro elementos da teoria dos fluidos (fogo, ar, água, terra), Galeno desenvolveu a teoria dos quatro temperamentos: melancólico (sujeito triste, reservado e sério), colérico (sujeito agressivo, excitável e impaciente), fleumático (sujeito passivo, pacífico e calmo) e sangüíneo (sujeito sociável, extrovertido e alegre). De fato, a origem mais sistemática dos testes psicológicos atuais pode ser traçada até o interesse dos estudiosos com a questão do número, desde a Idade Média, que iria desembocar nos psicólogos do fim do século passado na Alemanha, Inglaterra, França e Estados Unidos, onde se encontram as origens formais da psicometria. O estudo do número permitiu esta evolução por ser ele um símbolo que representa quantidade / extensão e que servia perfeitamente para basear técnicas de mensuração, que veio a ser a base da futura tecnologia de avaliação. Com isto fica igualmente acertado que a origem da avaliação formal e sistemática se deve aos psicólogos do fim do século passado.

De fato, estamos nos limiares do século XX. Em Psicologia vigoravam várias tendências epistemológicas, um tanto isoladas umas das outras, procurando superar o status pré-científico no estudo do psíquico (Boring, 1957).

A tradicional diatribe de origem cartesiana, alma vs. corpo, subsidiava estas tendências. Assim, temos, de um lado, a psicologia alemã da introspeção interessada na experiência subjetiva e, do outro, o empirismo inglês e norte-americano interessado no comportamento, bem como a escola (psicofísica) de Leipzig estudando os processos sensoriais. Estas duas grandes orientações se caracterizavam também pelo uso de procedimentos mais descritivos, no caso da psicologia introspectiva, e a procura de procedimentos mais quantitativistas por parte dos empiricistas. Portanto, não causa surpresa que as origens da Psicometria se encontrem no enfoque empiricista das psicologias da época. Desta sua origem, a Psicometria, tanto clássica quanto moderna (Teoria de Resposta ao Item - TRI) retém algumas caracterizações que permitem controvérsias. Entre elas, duas parecem particularmente fortes e, quiçá, preocupantes. Por um lado, a Psicometria, pelo menos na sua prática, é ainda guiada pela concepção positivista baconiana do empirismo (Bacon, 1620, 1984), isto é, que a ciência do universal se faz através do conhecimento do singular (indução), enfoque demonstrado como logicamente inviável, tanto pelo empiricista Hume (1739-1740) quanto por Popper (1972). Esta concepção creio ser responsável pelo descuido inaceitável da Psicometria com relação à teoria psicológica que deveria ser a preocupação preliminar e primordial na medida do psicológico. Preocupação esta que, felizmente, a Psicologia Cognitiva moderna procura ressuscitar. Por outro lado, predomina em Psicometria a concepção estatística sobre a psicológica. Os precursores e os que desenvolveram a Psicometria eram estatísticos de formação, tanto que ainda se define Psicometria como um ramo da Estatística, quando na verdade ela deve ser concebida como um ramo da Psicologia que interfaceia com a Estatística. Thurstone (1937) parecia preocupado com este problema, quando definiu como objeto de estudo para a sociedade psicométrica que acabara de fundar a "Psicologia Matemática", esta concebida como ramo da Psicometria dedicada à pesquisa dos modelos matemáticos dos processos psicológicos, mas sempre a serviço destes. De fato, a recente estatística (começou com Quetelet, 1796-1874) foi decisiva na elaboração da Psicometria e era entendido que através dela se podia avaliar adequadamente as diferenças individuais, objeto específico de qualquer avaliação. Temos neste contexto nomes e expoentes importantes tanto na Estatística quanto na Psicometria, como os de Sir Francis Galton (1822-1911), do seu discípulo Karl Pearson (1857-1936), Spearman (dos anos 1900) e Thurstone (dos anos 30). Esta plêiade de pesquisadores enviesou a Psicometria para o lado da Estatística, senão este que deve ser devidamente reparado no desenvolvimento daquela disciplina psicológica.

2. Origem da Psicometria

2.1 Apanhado histórico

A Psicometria (mais especificamente, os testes psicológicos) poderia ter tido origem em duas situações bastante distintas: 1) a psicologia de orientação empiricista ou 2) a psicologia mais mentalista de Binet na França. De fato, as duas tendências entraram em cena na mesma época para resolver os mesmos problemas, a saber, avaliar objetivamente as aptidões humanas. Apenas, Binet e Simon (1905) utilizando processos mentais e Galton (1883), Spearman (1904b) e outros empiricistas fazendo uso de processos comportamentais, mais especificamente, sensoriais. Embora o teste de inteligência de Binet tenha tido grande sucesso na Psicologia, não foi sua orientação que deu de fato a origem e o desenvolvimento à Psicometria porque lhe faltava o enfoque primordial da quantificação, que era o específico da orientação da psicologia empiricista. Da psicologia introspecionista da época realmente não se poderia esperar a origem da Psicometria, dada sua orientação puramente descritiva dos processos psicológicos. Conta Joncich (1968) que Thorndike (1904) ao enviar seu trabalho de medida em Psicologia a William James (que era da orientação descritiva) incluiu uma nota dizendo que o manuscrito era para seus alunos e que não aconselhava ao próprio James sua leitura!

Assim, a origem da Psicometria deve ser procurada nos trabalhos do estatístico Spearman (1904a, 1904b, 1907, 1913), que, no que se refere à Psicologia, seguiu os procedimentos fisicalistas de Galton (1883). Também não se deve estranhar que a Psicometria surgisse no campo das aptidões humanas (mentais, físicas, psicofísicas), pois, além de ser a temática psicológica da época, elas se coadunavam melhor a um estudo quantitativo, pois pode-se ali contabilizar o comportamento em termos de acertos e erros.

Aliás, para melhor entender a origem da Psicometria, pode-se seguir duas orientações, de início bastante independentes, que mais tarde se unificariam no que podemos chamar da Psicometria Clássica, a saber, a preocupação mais prática da

Psicometria e a preocupação mais teórica da Psicometria. A primeira tendência era mais visível em psicólogos com preocupações psico-pedagógicas e clínicas, cujo interesse nas provas psicológicas era detectar sobretudo o retardo mental e o potencial dos sujeitos para fins de predição na área acadêmica. A outra tendência visava mais o desenvolvimento da própria teoria psicométrica e era sobretudo perseguida por psicólogos de orientação estatística. Esta polaridade covaria com o que Boring (1957) chama de psicologia experimental e psicologia individual, esta mais preocupada com problemas humanos e aquela, mais com a ciência "pura". Este cisma seria somente superado lá pelos anos 1940, com a influência decisiva da orientação dos psicólogos da análise fatorial, especialmente de Thurstone (1938) com seus "Primary Mental Abilities". Estas tendências podem ser sumariamente visualizadas na exposição feita em 2.2.

A esta altura, parece relevante termos uma visão de conjunto do que aconteceu na história da Psicometria, desde sua origem até o presente momento, para, em seguida, desenvolver mais detalhadamente alguns temas desta história. Na verdade, seguindo Boring (1957), a história da avaliação psicológica tem sido, no início, dominada por alguns psicólogos expoentes em diferentes épocas. Assim, pode-se esquematizar esta história em termos da era de Galton, da era de Binet, etc., como veremos a seguir:

1) A *década de Galton*: 1880. Seus trabalhos visavam a avaliação das aptidões humanas através da medida sensorial, salientando-se sua obra "Inquiries into Human Faculty", de 1883. O trabalho de Galton terá enorme impacto tanto na orientação mais prática da psicometria (Cattell e outros psicometristas americanos), quanto na teórica (Pearson e Spearman).
2) A *década de Cattell*: 1890. Sob a influência da Galton, Cattell desenvolveu suas medidas das diferenças individuais e recolheu sua experiência no "Mental Tests and Measurements", de 1890, inaugurando, inclusive, a terminologia de "mental test".
3) A *década de Binet*: 1900. Foi a década onde predominaram os interesses da avaliação das aptidões humanas visando a predição na área acadêmica e na área da saúde. Embora Binet predomine de fato nesta época, outros expoentes aparecem neste período, salientando-se sobretudo Spearman na Inglaterra. Na verdade, no que se refere propriamente à teoria psicométrica, a década de 1900 deve ser considerada a *era de Spearman*, o qual deu os fundamentos da teoria da Psicometria clássica com suas obras "The proof and measurement of association between two things" (1904a), "'General intelligence' objectively determined and measured" (1904b), "Demonstration of formulae for true measurement of correlations" (1907) e "Correlations of sums and differences" (1913).
4) A era dos *testes de inteligência*: 1910 - 1930. Essa era se desenvolveu sob a influência de a) teste de inteligência de Binet-Simon (1905; 1908), b) artigo de Spearman sobre o fator G (1904b), c) a revisão do teste de Binet para os USA (Terman, 1916) e d) o impacto da primeira guerra mundial com a imposição da necessidade de seleção rápida, eficiente e universal de recrutas para o exército (os testes Army Alpha e Beta). Na verdade, temos uma série de psicólogos trabalhando nesta área, particularmente nos Estados Unidos. Goddard avaliava os imigrantes que entravam no país, vindo particularmente do sul e do leste europeu.

Cattell fundou um laboratório de Psicologia na Universidade da Pensilvânia, onde escreveu seu famoso artigo "Mental tests and their measurements", aparecido no Mind (1890); mais tarde fundou outro laboratório na Universidade de Columbia. Por outro lado, Joseph Jastrow desenvolveu 15 testes na Universidade de Wisconsin (1892-1927) e Hugo Müsterberg elaborou testes para crianças em seu laboratório em Harvard (1892-1916). Em 1904, um discípulo de Cattell, Edward L. Thorndike, publicou o primeiro livro tratando de medidas mentais e educacionais e seu discípulo, Cliff W. Stone, publicou o primeiro teste padronizado em aritmética, o Stone Arithmetic Test (1908). Com a advento da I Guerra Mundial, vários psicólogos (Yerkes, Thorndike, Seashore, Angell) desenvolveram testes de seleção para soldados. Nos anos 20 houve uma avalanche de novos testes, ultrapassando as centenas deles, dos quais muito poucos resistiram até o presente.

5) A década da *análise fatorial*: 1930. Já por volta de 1920, o entusiasmo com os testes de inteligência vinha caindo muito, sobretudo quando se mostrou que eles eram demasiadamente dependentes da cultura onde eram criados, não apoiando a idéia de um fator geral universal do estilo Spearman. Tais eventos fizeram com que os psicólogos estatísticos começassem a repensar as idéias de Spearman. De fato, Kelley quebrou com a tradição de Spearman em 1928. Esta tendência foi seguida, na Inglaterra, por Thomson (1940) e Burt (1941) e nos USA, por Thurstone (1935, 1947). Este autor é especialmente relevante nesta época, pois além de desenvolver a análise fatorial múltipla, atuou no desenvolvimento da escalagem psicológica (1927, 1928, Thurstone & Chave, 1929), bem como fundando, em 1936, a Sociedade Psicométrica Americana, juntamente com a revista Psychometrika, ambas dedicadas ao estudo e avanço da Psicometria.
6) A *era da sistematização*: 1940 - 1980. Esta época é marcada por duas tendências opostas: os trabalhos de síntese e os de crítica. Nas obras de síntese, temos Guilford (1936, "Psychometric Methods", reeditada em 1954), tentando sistematizar os avanços em Psicometria até então conseguidos; Gulliksen (1950, "Theory of Mental Tests"), sistematizando a teoria clássica dos testes psicológicos e Torgerson (1958, "Theory and Methods of Scaling"), sistematizando a teoria sobre a medida escalar. Além disso, Thurstone (1947) e Harman (1967) recolheram os avanços na área da análise fatorial; Cattell (1965; Cattell & Warburton, 1967) procurou sintetizar os dados da medida em personalidade e Guilford (1967) procurou sistematizar uma teoria sobre a inteligência. Por outro lado, Buros (1938) iniciou uma coletânea de todos os testes existentes no mercado, a qual vem sendo refeita periodicamente (mais ou menos a cada cinco anos), o "Mental Measurement Yearbook". Na mesma época, A "American Psychological Association" - APA (1954, 1974, 1985) introduziu as normas de elaboração e uso dos testes.

No lado da crítica, temos Stevens (1946) levantando o problema das escalas de medida que deu/dá muita polêmica na área (Lord, 1953; Gaito, 1980; Michell, 1986; Townsend & Ashby, 1984) e, sobretudo, surge a primeira grande crítica à teoria clássica dos testes na obra de Lord e Novick (1968 - "Statistical Theory of Mental Tests Scores") que iniciou o desenvolvimento de uma teoria alternativa, a teoria do traço latente, que vai desembocar na teoria moderna da Psicometria, a Teoria de Resposta ao Item – TRI, mais tarde sintetizada por Lord (1980). Outra tendên-

cia de crítica para superar as dificuldades da Psicometria clássica foi iniciada pela Psicologia Cognitiva de Sternberg (1977,1982, 1985; Sternberg & Detterman, 1979; Sternberg & Weil, 1980) com seu modelo, procedimentos e pesquisas sobre os componentes cognitivos, na área da inteligência.

7) A era da *Psicometria moderna* (Teoria de Resposta ao Item – TRI): 1980. Chamar a era atual de era da TRI talvez seja inadequado, porque 1) esta teoria, embora esteja sendo o modelo no dito primeiro mundo, ainda não resolveu todos seus problemas fundamentais para se tornar o modelo moderno definitivo de Psicometria e 2) ela não veio para substituir toda a Psicometria clássica, mas apenas partes dela. De qualquer forma é o que há de mais novo no campo. Aliás, poderíamos sintetizar melhor o que está ocorrendo hoje no mundo da Psicometria, arrolando três/quatro linhas genéricas em que os psicometristas vêm atuando:
   a) Sistematização da Psicometria Clássica: Anastasi (1988), Crocker e Algina, 1986; Thorndike, 1982.
   b) Sistematização e pesquisa na TRI: Lord (1980), Hambleton e Swaminathan (1985), Hambleton, Swaminathan e Rogers (1991) sistematizam esta área e mostram a quantidade de pesquisa que nela está sendo realizada.
   c) Pesquisa em uma série de áreas paralelas da Psicometria:
      - testes com referência a critério (Berk, 1984)
      - testes sob medida ("computer adaptive testing" - Wainer, 1990)
      - banco de itens ("Applied Psychological Measurement", 1987; Millman & Arter, 1984; Wright & Bell, 1984)
      - equiparação dos escores (Angoff, 1984; Holland & Rubin, 1982; Skaggs & Lissitz, 1986)
      - validade dos testes (Wainer & Braun, 1988)
      - vieses dos testes (Berk, 1982; Reynolds & Brown, 1984; Osterling, 1983)
      - construção de itens (Brown, 1983; Gronlund, 1988; Mehrens & Lemann, 1984; Osterling, 1989; Roid, 1984; Roid & Haladyna, 1980, 1982).
   d) Neste contexto podemos igualmente situar o impacto dos trabalhos da Psicologia Cognitiva (Sternberg, 1977, 1982, 1985; Sternberg & Detterman, 1979; Sternberg & Weil, 1980; Carpenter, Just, & Shell, 1990) com suas pesquisas na área das aptidões através do estudo dos componentes cognitivos.
   e) Finalmente, vale a pena relacionar as principais revistas onde estão sendo hoje publicados os trabalhos de Psicometria (em parênteses, o ano de fundação da revista):
      - Psychometrika (1936)
      - Educational and Psychological Measurement (1941)
      - The British Journal of Mathematical and Statistical Psychology (1948)
      - Journal of Educational Measurement (1964)
      - Journal of Educational Statistics (1976)
      - Applied Psychological Measurement (1977)
      - Psychological Bulletin (1903)
      - Behavior Reseach Methods, Instruments, & Computers (1969).

2.2 Os testes psicológicos

Os testes psicológicos que iam surgindo no final do século passado e nas primeiras décadas deste representaram o campo propício onde a Psicometria se originou e mais se desenvolveu. Assim, algumas notas históricas neste campo são úteis para estudar o desenvolvimento da própria teoria psicométrica.

Embora haja relatos de uso de testes para seleção de funcionários civis da China lá por 3.000 A.C. (Dubois, 1970), as origens efetivas destes instrumentos psicológicos podem ser traçadas aos trabalhos de Galton (1822-1911) no seu laboratório de Kensington, Inglaterra.

De fato, haviam dois tipos de preocupações na área da avaliação do psicológico:

1) Preocupação psico-pedagógica e psiquiátrica na França (Esquirol, Seguin, Binet). Esta tendência se preocupava com o tratamento mais humano a ser dado aos doentes mentais que eram definidos por retardos mentais mais ou menos graves, havendo, portanto, diferentes níveis de doença mental ou retardo mental. É o trabalho do médico francês Esquirol (1838). De interesse para a Psicometria é sua preocupação com a questão de como identificar o nível de retardo mental. Concluiu ele que é na área da linguagem (uso da língua) onde estaria o critério para tal decisão. Seu colega Seguin (1866-1907) também se preocupou com o retardo mental, mas sua atuação foi mais no sentido de tratar esses deficientes através de treinamento fisiológico. Na mesma linha de ação se encontra outro francês, o psicólogo Binet que desenvolveu um teste mental para avaliar o retardo mental (sobre ele, mais adiante).
2) Preocupação experimentalista (Alemanha, Inglaterra e USA). A preocupação central dos psicólogos desta orientação era a descoberta de uniformidades no comportamento dos indivíduos, não tanto as diferenças individuais (como na escola francesa). Aliás, as diferenças eram concebidas como desvios ou erros. Seus temas caíam sobre o comportamento sensorial, preocupação que espelha a origem destes psicólogos como físicos e fisiólogos. Um outro elemento importante para a futura psicometria foi a preocupação com o controle das condições em que se faziam as observações. Um enfoque mais individual neste grupo de psicólogos foi o de Cattell, psicólogo americano estudando na Europa, que se interessou sobretudo precisamente pelas diferenças individuais dos sujeitos (dele, mais adiante).

Alguns expoentes destas tendências serão brevemente detalhados a seguir.

Galton (1883) acreditava que as operações intelectuais poderiam ser avaliadas através de medidas sensoriais. Dado que, dizia ele, toda a informação do homem chega pelos sentidos, quanto melhor o estado destes, melhores seriam as operações intelectuais. Assim, ele se preocupou em estabelecer os parâmetros das dimensões ideais dos sentidos, fazendo um levantamento amplo de medidas sensoriais. Ele considerava particularmente importante nos indivíduos a capacidade de discriminação sensorial do tato e dos sons. Galton de fato contribuiu para a Psicometria em três áreas: 1) medida da discriminação sensorial, onde desenvolveu testes, cujos conceitos são ainda utilizados (barras para medir percepção de comprimento, apito

para percepção de altura do tom); 2) escalas de pontos, questionários e associação livre, que ele utilizava após as medidas sensoriais; 3) desenvolvimento e simplificação de métodos estatísticos para analisar quantitativamente os dados coletados, tarefa levada adiante pelo seu depois famoso discípulo Karl Pearson.

James McKeen Cattell, psicólogo americano, fez sua tese em Leipzig sobre diferenças individuais no tempo de reação, apesar do seu orientador e estudioso do mesmo tema Wundt não gostar deste tipo de pesquisa, dado que este estava a procura de uniformidades e não de diferenças individuais. Mais tarde, como professor em Cambridge (1988) ficou mais animado com a sua orientação vendo e sentindo a influência de Galton que também trabalhava com a medida das diferenças individuais. Famoso é seu artigo de 1890, porque nele Cattell usa pela primeira vez a expressão, que fez sucesso internacional e histórico, de teste mental ("mental test") para as provas aplicadas anualmente aos alunos universitários no sentido de avaliar seu nível intelectual nos USA. Cattell seguiu as idéias de Galton, dando ênfase às medidas sensoriais porque elas permitiam maior precisão. Percebeu ele que medidas objetivas para funções mais complexas, que vinham sendo usadas sobretudo na Alemanha, tais como testes contendo operações simples de aritmética, testes de memória e resistência à fadiga (Kraepelin, 1895), bem como testes de cálculo, duração de memória e complementação de sentenças (Ebbinghaus, 1897), não produziam resultados condizentes com o desempenho acadêmico. Contudo, os próprios testes de Cattell também não produziam resultados congruentes entre si (Sharp, 1899; Wissler, 1901) e nem correlacionavam com a avaliação que os professores faziam do nível intelectual dos alunos (Bolton, 1892; Gilbert, 1894) e nem mesmo correspondiam ao desempenho acadêmico desses alunos (Wissler, 1901).

Binet e *Henri* (1896) começaram com uma séria crítica a todos estes testes, afirmando que eles 1) ou eram puramente medidas sensoriais que, embora permitindo maior precisão, não tinham relação importante com as funções intelectuais (irrelevância) ou 2) se eram testes de conteúdo intelectual, estes se dirigiam a habilidades demasiadamente específicas, como o puro memorizar, calcular, etc., quando os testes deveriam se orientar para medir funções mais amplas como a memória, imaginação, atenção, compreensão etc. De fato, *Binet* e *Simon* (1905) desenvolveram seu famoso teste de 30 itens para cobrir uma gama variada de funções (como julgamento, compreensão e raciocínio) com o objetivo de avaliar o nível de inteligência de crianças e adultos, através do qual estavam especialmente interessados em detectar o retardo mental. Esta orientação de Binet e Simon em elaborar testes de conteúdo mais cognitivo (e não sensorial) e cobrindo funções mais amplas (não específicas) fez grande sucesso nos anos subseqüentes, especialmente nos USA com a tradução do seu teste por Terman (1916), inaugurando de vez a era dos testes, inclusive com a introdução do Q.I., sendo

$$Q.I. = 100\ (IM / IC)$$

onde,

Q.I. = quociente intelectual
IM = idade mental
IC = idade cronológica

Este quociente substituiu a forma de Binet e Simon de expressar o nível intelectual do sujeito em termos de Idade Mental ("Age Mentale", a saber, a criança teria aquela idade mental se respondesse corretamente as questões que tipicamente crianças de tal idade cronológica eram capazes de responder corretamente).

Após estes primórdios, os testes se popularizam, sobretudo com a vinda da Primeira Guerra Mundial, na qual o exército americano desenvolveu uma série de baterias de testes (*Army Alpha* e *Army Beta*) para seleção de soldados, introduzindo, inclusive, os testes de aplicação coletiva (até o momento, os testes eram todos de aplicação individual). Finda a guerra, a indústria e as instituições em geral iniciaram o uso massiço dos testes. No campo das aptidões, contudo, foi *Thurstone* (1938, 1941) quem deu impulso inovador a estas técnicas com o uso da análise fatorial, da qual foi um expoente teórico, e sua bateria "Primary Mental Abilities", que incentivou o aparecimento de uma plêiade de outras baterias (DAT, PMA, GATB, TEA, WISC, WAIS). A área da personalidade não ficou atrás. Testes e inventários de personalidade surgiam às dezenas (MMPI, 16PF, EPPS, POI, CPI, CEP, EPI), além de instrumentais menos objetivos, os ditos testes projetivos (TAT, CAT, Rosenzweig, Szondi, Rorschach, HTP). Estava, enfim, instalada a tecnocracia dos testes e da Psicometria.

Bibliografia

American Psychological Association (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: American Psychological Association.

American Psychological Association (1974, 1985). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Anastasi, A. (1988). *Psychological testing*. 6th ed. New York: Macmillan Pub. Co.

Angoff, W.H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.

Angoff, W.H. & Ford, S.F. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*, 95-105.

Applied Psychological Measurement (1987). Special series: Problems, perspectives, and practical issues in equating, *11*(3).

Bacon, F. (1620; 1984). *Novum Organum*. In V. Civita, Os Pensadores: Francis Bacon, 1984. São Paulo: Abril S.A.

Barclay, J.R. (1991). *Psychological assessment: A theory and systems approach*. Malabar, FL: Krieger Publishing Company.

Berk, R.A. (Ed. - 1982). *Handbook of methods for detecting test bias*. Baltimore, MD: The Johns Hopkins University Press.

Berk, R.A. (Ed. - 1984). *A guide to criterion-referenced test construction*. Baltimore, MD: The Johns Hopkins University Press.

Binet, A. & Henri, V. (1896). La psychologie individuelle. *L'Année Psychologique, 2*, 411-465.

Binet, A. & Simon, Th. (1908). Le développement de l'intelligence chez les enfants. *Année Psychologique, 14*, 1-94.

Binet, A. & Simon. Th. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique, 11*, 191-244.

Bolton, T.L. (1892). Growth of memory in school children. *American Journal of Psychology, 4*, 362-380.

Boring, E.G. (1957). *A history of experimental psychology*. 2d ed. New York: Appleton-Century-Crofts, Inc.

Brown, F.G. (1983). *Principles of education and psychology testing*. New York: Holt, Rinehart, and Winston.

Buros, O.K. (Ed. - 1938). *The first mental measurement yearbook*. Highland Park: Gryphon Press.

Burt, C. (1941). *The factors of the mind*. New York: Macmillan.

Carpenter, P.A., Just, M.A., & Shell, P. (1990). What one intelligence test measures: A theoretical account on the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*(3), 404-431.

Cattell, J.M. (1890). Mental tests and measurements. *Mind, 15*, 373-380.

Cattell, R.B. & Warburton, F.W. (1967). *Objective personality and motivation tests*. Urbana, IL: University of Illinois Press.

Cattell, R.B. (1965). *The scientific analysis of personality*. Baltimore, MD: Penguin Books, Inc.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.

Dubois, P.H. (1970). *A history of psychological testing*. Boston: Allyn and Bacon.

Ebbinghaus, H. (1897). Ueber eine neue Methode zur Prüfung geistiger Fähigkeiten und ihre Anwendung bei Schulkindern. *Zsch. Psychol., 13*, 401-459.

Esquirol, J.E.D. (1838). *Des maladies mentales considérées sous les rapports médical, hygiénique et médico-légal* (2 vols.). Paris: Baillière.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin, 87*, 564-567.

Galton, F. (1883). *Inquiries into human faculty and its development*. London: Macmillan.

Gilbert, J.A. (1894). Researches on mental and physical development of schoolchildren. *Studies from Yale Psychol. Lab, 2*, 40-100.

Gronlund, N.E. (1988). *How to construct achievement tests*, 4th ed. Englewood Cliffs, NJ: Prentice-Hall.

Guilford, J.P. (1936, 1954). *Psychometric methods*. New York: McGraw-Hill.

Guilford, J.P. (1959). *Personality*. New York: McGraw-Hill.

Guilford, J.P. (1967). *The nature of human intelligence*. New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Gulliksen, H. (1987). *Theory of mental tests*. Hillsdale, NJ: Erlbaum.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Norwell, MA: Kluwer Nijhoff.

Hambleton, R.K., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of item response theory*. Beverly Hills, CA: SAGE.

Harman, H.H. (1967). *Modern factor analysis*. Chicago, IL: University of Chicago Press.

Holland, P.W. & Rubin, D.B. (Eds. - 1982). *Test equating*. Orlando, FL: Academic Press.

Hume, D. (1739-40). *Treatise of human nature*, vol. I Editado por L.A. Selby-Bigge em 1888. Oxford: Clarendon Press.

Joncinch, G. (1968). *The sane positivist: A biography of Edward L. Thorndike*. Middletown: Wesleyan University Press.

Kelley, T.L. (1928). *Crossroads in the mind of man*. Stanford, CA: Stanford University Press.

Kraepelin, E. (1895). Der psychologische Versuch in der Psychiatrie. *Psychol. Arbeiten, 1*, 1-91.

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental tests scores*. Reading, MA: Addison-Wesley.

Lord, F.M. (1953). On the statistical treatment of football numbers. *The American Psychologist, 8*, 750-751.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Mehrens, W.A. & Lemann, I.J. (1984). *Measurement and evaluation in education and psychology*, 3d ed. New York: Holt, Rinehart, & Winston.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin, 100*, 398-407.

Millman, J. & Arter, J.A. (1984). Issues in item banking. *Journal of Educational Measurement, 21*, 315-330.

Osterling, S.J. (1983). *Test item bias*. Beverly Hills, CA: SAGE.

Osterling, S.J. (1989). *Constructing test items*. Boston: Kluwer Academic Publs.

Popper, K. R. (1972). *A lógica da pesquisa científica*. São Paulo: Editora Cultrix.

Reynolds, C.R. & Brown, R.T. (Eds. - 1984). *Perspectives on bias in mental testing*. New York: Plenum Press.

Roid, G.H. & Haladyna, T.M. (1980). The emergence of an item-writing technology. *Review of Educational Research, 50*, 293-314.

Roid, G.H. & Haladyna, T.M. (1982). *A technology for test item writing*. New York: Academic Press.

Roid, G.H. (1984). Generating the test items. In R.A. Berk (ed.), *A guide to criterion-referenced test construction*. Baltimore: Johns Hopkins University Press, 49-77.

Sharp, S.E. (1899). Individual psychology: Study in psychological methods. *American Journal of Psychology, 10,* 329-391.

Skaggs, G. & Lissitz, R.W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56,* 495-529.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology,* 18, 161-169.

Spearman, C. (1904a). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72-101.

Spearman, C. (1904b). 'General intelligence' objectively determined and measured. *American Journal of Psychology, 15,* 201-293.

Spearman, C. (1913). Correlations of sums and differences. *British Journal of Psychology, 5,* 417-426.

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

Sternberg, R.J. & Detterman, D.K. (eds. - 1979). *Human intelligence: Perspectives on its theory and measurement.* Norwood, NJ: Ablex.

Sternberg, R.J. & Rifkin, B. (1979). The development of analogical reasoning processes. *Journal of Experimental Child Psychology, 27,* 195-232.

Sternberg, R.J. & Weil, E.M. (1980). An aptitude x strategy interaction in linear syllogistic reasoning. *Journal of Educational Psychology, 72,* 226-239.

Sternberg, R.J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities.* Hillsdale, NJ: Erlbaum.

Sternberg, R.J. (1980). Representation and process in linear syllogistic reasoning. *Journal of Experimental Psychology, General, 109,* 119-159.

Sternberg, R.J. (1984). What cognitive psychology can (and cannot) do for test development. In B.S. Plake (Ed.), *Social and technical issues in testing: Implications for test construction and usage.* Hillsdale, NJ: Erlbaum, 39-60.

Sternberg, R.J. (1985). General intellectual ability. In R.J. Sternberg (ed.), *Human abilities: An information-processing approach.* New York: Freeman, 5-29.

Sternberg, R.J. (1990). T & T is an explosive combination: Technology and testing. *Educational Psychologist, 25,* 201-222.

Sternberg, R.J. (ed. - 1982). *Advances in the psychology of human intelligence.* Hillsdale, NJ: Erlbaum.

Sternberg. R.J. (1979). The nature of mental abilities. *American Psychologist, 34,* 214-230.

Stevens, S.S (1946). On the theory of scales of measurement. *Science, 103,* 667-680.

Stone, C.W. (1908). Stone Arithmetic Test. Em E.G. Boring, A hisory fo experimental psychology (1957). Englewood Cliffs, NJ: Prentice-Hall.

Sundberg, N.D. (1977). *Assessment of persons.* Englewood Cliffs: Prentice-Hall.

Terman, L.M. (1916). The measurement of intelligence. Boston, MA: Houghton Mifflin.

Thomson, G.H. (1940). Weighting for battery reliability and prediction. *British Journal of Psychology, 30,* 357-366.

Thorndike, E.L. (1904). *An introduction to the theory of mental and social measurements.* New York: Science Press.

Thorndike, R.L. (1982). *Applied psychometrics.* Boston: Hougton Mifflin.

Thurstone, L.L. (1952). The criterion problem in personality research. *Psychometric Lab. Rep.,* No. 78. Chicago, IL: University of Chicago.

Thurstone, L.L (1927). A law of comparative judgment. *Psychological Review, 34,* 273-286.

Thurstone, L.L (1928). Attitudes can be measured. *American Journal of Sociology, 33,* 529-554.

Thurstone, L.L (1937). Psychology as a quantitative rational science. *Science, 85,* 227-232.

Thurstone, L.L & Chave, E.J. (1929).*The measurement of attitudes.* Chicago, IL: University of Chicago Press.

Thurstone, L.L. & Thurstone, T.G. (1941). Factorial studies of intelligence. *Psychometric Monographs,* no. 2.

Thurstone, L.L. (1931). *The reliability and validity of tests.* Ann Arbor, MI: Edward Brothers.

Thurstone, L.L. (1935, 1947). *Multiple factor analysis.* Chicago, IL: University of Chicago Press (publicado em 1935 como "Vectors of the mind").

Thurstone, L.L. (1938). Primary mental abilities. *Psychometric Monographs,* n. 1.

Torgerson, W.S. (1958). *Theory and methods of scaling.* New York: Wiley.

Townsend, J.T. & Ashby, F.G. (1984) Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin, 96,* 394-401.

Wainer, H. & Braun, H.I. (Eds. - 1988). *Test validity.* Hillsdale, NJ: LEA.

Wainer, H. (1983). Are we correcting for guessing in the wrong direction? In D.J. Weis (ed.), *New horizons in testing.* New York: Academic Press.

Wainer, H. (Ed. - 1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: LEA.

Wissler, C. (1901). Correlation of mental and physical traits. *Psychological Monographs, 3,* no. 16.

Wright, B.D. & Bell, S.R. (1984). Items banks: What, why, how. *Journal of Educational Measurement, 21,* 331-346.

TAXONOMIA DOS INSTRUMENTOS PSICOLÓGICOS

Luiz Pasquali

Classificar a diversidade enorme de tipos de testes psicológicos se constitui presentemente em uma tarefa muito arbitrária. Muitos diferentes critérios podem ser utilizados para tal tipologia. Na verdade, cada autor tem a sua classificação, dependendo do ponto de vista utilizado ou do gosto pessoal. Uns classificam-nos em termos dos traços latentes que estes mesmos testes medem, como Kline (1993) que os divide em: testes de inteligência, testes de aptidões, testes de habilidades e desempenho, questionários de personalidade, testes objetivos e projetivos de personalidade, testes de interesses e motivação, outros tipos de testes psicológicos. Anastasi (1988), por sua vez, divide os testes em: testes de inteligência geral, testes de aptidões específicas e testes de personalidade, incluindo dentro destas três divisões uma série de outras subdivisões.

Esta é uma situação infeliz, porque uma boa taxonomia se apresenta como necessária para poder se pôr alguma ordem dentro de um sistema, o qual permitiria mapear a campo dos testes, visualizar em que área os testes abundam e em quais eles fazem falta, o que evidentemente seria de capital importância para o desenvolvimento na área dos testes. Uma classificação em qualquer área de saber parece necessária para um conhecimento sistemático da mesma. Na verdade, já em 1967, Cattell e Warburton tentavam estabelecer uma lógica para uma taxonomia racional dos testes psicológicos. Embora reconhecendo a importância que tem uma taxonomia em qualquer ciência, em Psicologia tal taxonomia simplesmente não existia. Assim, eles mesmos empreenderam a tarefa para viabilizar uma tal classificação dos testes psicológicos. Dois princípios de taxonomia, diziam eles, poderiam ser utilizados, a saber, classificar os testes em termos dos traços latentes que eles medem ou classificá-los em cima das operações concretas utilizadas na construção dos testes. Acharam eles que o último critério seria mais útil, pois o número possível de traços latentes parece quase ilimitado, tornando a tarefa de classificação sem muito sentido e utilidade, se feita em termos dos traços latentes. Este critério, isto é, basear a classificação em aspectos concretos dos testes, comporta uma divisão tripartite destes, a saber, em termos do tipo de instruções que define o modo de resposta, da situação estímulo do teste e do modo de apurar a resposta ao teste; dentro desta divisão uma centena de outras seriam possíveis. Na verdade, esta tentativa de Cattell não teve nenhum impacto no desenvolvimento de uma taxonomia sistematizada na área dos testes, que ainda continua ao bel-prazer de cada autor.

Sharp, S.E. (1899). Individual psychology: Study in psychological methods. *American Journal of Psychology, 10*, 329-391.

Skaggs, G. & Lissitz, R.W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56*, 495-529.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology,* 18, 161-169.

Spearman, C. (1904a). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72-101.

Spearman, C. (1904b). 'General intelligence' objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

Spearman, C. (1913). Correlations of sums and differences. *British Journal of Psychology, 5*, 417-426.

Spearman, C. (1927). *The abilities of man*. New York: Macmillan.

Sternberg, R.J. & Detterman, D.K. (eds. - 1979). *Human intelligence: Perspectives on its theory and measurement*. Norwood, NJ: Ablex.

Sternberg, R.J. & Rifkin, B. (1979). The development of analogical reasoning processes. *Journal of Experimental Child Psychology, 27*, 195-232.

Sternberg, R.J. & Weil, E.M. (1980). An aptitude x strategy interaction in linear syllogistic reasoning. *Journal of Educational Psychology, 72*, 226-239.

Sternberg, R.J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. Hillsdale, NJ: Erlbaum.

Sternberg, R.J. (1980). Representation and process in linear syllogistic reasoning. *Journal of Experimental Psychology, General, 109*, 119-159.

Sternberg, R.J. (1984). What cognitive psychology can (and cannot) do for test development. In B.S. Plake (Ed.), *Social and technical issues in testing: Implications for test construction and usage*. Hillsdale, NJ: Erlbaum, 39-60.

Sternberg, R.J. (1985). General intellectual ability. In R.J. Sternberg (ed.), *Human abilities: An information-processing approach*. New York: Freeman, 5-29.

Sternberg, R.J. (1990). T & T is an explosive combination: Technology and testing. *Educational Psychologist, 25*, 201-222.

Sternberg, R.J. (ed. - 1982). *Advances in the psychology of human intelligence*. Hillsdale, NJ: Erlbaum.

Sternberg. R.J. (1979). The nature of mental abilities. *American Psychologist, 34*, 214-230.

Stevens, S.S (1946). On the theory of scales of measurement. *Science, 103*, 667-680.

Stone, C.W. (1908). Stone Arithmetic Test. Em E.G. Boring, A hisory fo experimental psychology (1957). Englewood Cliffs, NJ: Prentice-Hall.

Sundberg, N.D. (1977). *Assessment of persons*. Englewood Cliffs: Prentice-Hall.

Terman, L.M. (1916). The measurement of intelligence. Boston, MA: Houghton Mifflin.

Thomson, G.H. (1940). Weighting for battery reliability and prediction. *British Journal of Psychology, 30*, 357-366.

Thorndike, E.L. (1904). *An introduction to the theory of mental and social measurements*. New York: Science Press.

Thorndike, R.L. (1982). *Applied psychometrics*. Boston: Hougton Mifflin.

Thurstone, L.L. (1952). The criterion problem in personality research. *Psychometric Lab. Rep.*, No. 78. Chicago, IL: University of Chicago.

Thurstone, L.L (1927). A law of comparative judgment. *Psychological Review, 34*, 273-286.

Thurstone, L.L (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529-554.

Thurstone, L.L (1937). Psychology as a quantitative rational science. *Science, 85*, 227-232.

Thurstone, L.L & Chave, E.J. (1929).*The measurement of attitudes*. Chicago, IL: University of Chicago Press.

Thurstone, L.L. & Thurstone, T.G. (1941). Factorial studies of intelligence. *Psychometric Monographs*, no. 2.

Thurstone, L.L. (1931). *The reliability and validity of tests*. Ann Arbor, MI: Edward Brothers.

Thurstone, L.L. (1935, 1947). *Multiple factor analysis*. Chicago, IL: University of Chicago Press (publicado em 1935 como "Vectors of the mind").

Thurstone, L.L. (1938). Primary mental abilities. *Psychometric Monographs*, n. 1.

Torgerson, W.S. (1958). *Theory and methods of scaling*. New York: Wiley.

Townsend, J.T. & Ashby, F.G. (1984) Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin, 96*, 394-401.

Wainer, H. & Braun, H.I. (Eds. - 1988). *Test validity*. Hillsdale, NJ: LEA.

Wainer, H. (1983). Are we correcting for guessing in the wrong direction? In D.J. Weis (ed.), *New horizons in testing*. New York: Academic Press.

Wainer, H. (Ed. - 1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: LEA.

Wissler, C. (1901). Correlation of mental and physical traits. *Psychological Monographs, 3*, no. 16.

Wright, B.D. & Bell, S.R. (1984). Items banks: What, why. how. *Journal of Educational Measurement, 21*, 331-346.

TAXONOMIA DOS INSTRUMENTOS PSICOLÓGICOS

Luiz Pasquali

Classificar a diversidade enorme de tipos de testes psicológicos se constitui presentemente em uma tarefa muito arbitrária. Muitos diferentes critérios podem ser utilizados para tal tipologia. Na verdade, cada autor tem a sua classificação, dependendo do ponto de vista utilizado ou do gosto pessoal. Uns classificam-nos em termos dos traços latentes que estes mesmos testes medem, como Kline (1993) que os divide em: testes de inteligência, testes de aptidões, testes de habilidades e desempenho, questionários de personalidade, testes objetivos e projetivos de personalidade, testes de interesses e motivação, outros tipos de testes psicológicos. Anastasi (1988), por sua vez, divide os testes em: testes de inteligência geral, testes de aptidões específicas e testes de personalidade, incluindo dentro destas três divisões uma série de outras subdivisões.

Esta é uma situação infeliz, porque uma boa taxonomia se apresenta como necessária para poder se pôr alguma ordem dentro de um sistema, o qual permitiria mapear a campo dos testes, visualizar em que área os testes abundam e em quais eles fazem falta, o que evidentemente seria de capital importância para o desenvolvimento na área dos testes. Uma classificação em qualquer área de saber parece necessária para um conhecimento sistemático da mesma. Na verdade, já em 1967, Cattell e Warburton tentavam estabelecer uma lógica para uma taxonomia racional dos testes psicológicos. Embora reconhecendo a importância que tem uma taxonomia em qualquer ciência, em Psicologia tal taxonomia simplesmente não existia. Assim, eles mesmos empreenderam a tarefa para viabilizar uma tal classificação dos testes psicológicos. Dois princípios de taxonomia, diziam eles, poderiam ser utilizados, a saber, classificar os testes em termos dos traços latentes que eles medem ou classificá-los em cima das operações concretas utilizadas na construção dos testes. Acharam eles que o último critério seria mais útil, pois o número possível de traços latentes parece quase ilimitado, tornando a tarefa de classificação sem muito sentido e utilidade, se feita em termos dos traços latentes. Este critério, isto é, basear a classificação em aspectos concretos dos testes, comporta uma divisão tripartite destes, a saber, em termos do tipo de instruções que define o modo de resposta, da situação estímulo do teste e do modo de apurar a resposta ao teste; dentro desta divisão uma centena de outras seriam possíveis. Na verdade, esta tentativa de Cattell não teve nenhum impacto no desenvolvimento de uma taxonomia sistematizada na área dos testes, que ainda continua ao bel-prazer de cada autor.

Antes mesmo da tentativa de Cattell e Warburton, Campbell (1957) também tentara uma tipologia para os testes, dividindo-os em três dicotomias, a saber:

- voluntário vs. objetivo: conforme existir ou não uma resposta correta no teste;
- direto vs. indireto: conforme o objetivo do teste ser óbvio ou disfarçado;
- resposta livre vs. resposta estruturada: conforme existir ou não uma alternativa a ser marcada.

Combinando estas dicotomias, surgem oitos tipos de testes:

- voluntário, indireto, resposta livre: testes projetivos;
- voluntário, indireto, resposta estruturada: testes projetivos com resposta a base de escolha de alternativas oferecidas, técnica Q-sort;
- voluntário, direto, resposta livre: complementação de sentenças, relato;
- voluntário, direto, resposta estruturada: escala tipo Thurstone ou Likert, Kuder, MMPI, inventários de personalidade;
- objetivo, indireto, resposta livre: "verbal summator", "Intuition Questionaire" (Sherriffs, 1948), "Autokinetic Word Technique" (Rechtschaffen & Mednick, 1955);
- objetivo, indireto, resposta estruturada: os anteriores com resposta estruturada;
- objetivo, direto, resposta livre: testes de aptidão com resposta livre;
- objetivo, direto, resposta estruturada: testes de aptidão de múltipla escolha.

Esta tentativa de Campbell também não deu em nada. Diante desta situação, não será tentada nenhuma fundamentação teórica sobre a taxonomia dos testes neste capítulo. Entretanto, para pôr alguma ordem na casa, devemos apresentar uma classificação dos instrumentos psicológicos, sobretudo em vista ao fato de que diferentes testes têm diferentes técnicas e teorias que fundamentam sua construção. Por isso, propomos a seguinte taxonomia, simplesmente porque estes vários tipos de instrumentos apresentam técnicas diferentes tanto de construção quanto de aferição dos seus parâmetros de validade e fidedignidade:

1. Testes referentes a critério
2. Testes referentes a construto
3. Testes referentes a conteúdo
4. Testes comportamentais: Observação do comportamento
5. Levantamentos: "survey"
6. Novas tecnologias.

1. Testes Referentes a Critério

1.1 Conceituação

Estes instrumentos, como o nome diz, se caracterizam por serem capazes de discriminar grupos-critério. Isto quer dizer que tais testes são construídos para diferenciar grupos distintos naquilo que o teste pretende medir e adquirem sua validade pela capacidade de poderem ou não diferenciar claramente estes grupos-critério. Por exemplo, o MMPI foi construído a partir de grupos-critério psiquiátricos. A lógica que fundamentou a construção deste teste foi a de que o conjunto de tarefas (itens, comportamentos) que fossem respondidos de uma maneira típica por

um grupo psiquiátrico, diferentemente de todos os outros grupos psiquiátricos e dos sujeitos normais, seria um teste que definiria este grupo. Assim, o grupo dos psicopatas respondeu a uma série de itens (entre as várias centenas que os autores coletaram) de uma maneira diferente dos sujeitos normais e de todos os outros tipos de sujeitos psiquiátricos (neuróticos, esquizofrênicos, etc.). Consequentemente, esta série de comportamentos constituem o teste que mede psicopatia. E assim foi feito com as demais categorias psiquiátricas, em número de 10.

1.2 Aplicações

Estes testes são úteis e são utilizados quando se quer precisamente discriminar sujeitos em termos de pertença a uma ou outra classe e são, por isso, de uso corrente em psicologia aplicada, como em seleção, diagnóstico psiquiátrico e orientação acadêmica e vocacional.

Em *seleção*, por exemplo, se se deseja eliminar certo tipo de pessoas, como os demasiadamente ansiosos numa seleção para pilotos de caça, constrói-se um teste composto de uma série de comportamentos que distingue claramente os ansiosos dos não ansiosos. No caso do *diagnóstico psiquiátrico*, se se deseja definir em que grupo psiquiátrico o sujeito cai, constrói-se um teste composto de várias séries de itens que discriminam os psicopatas dos sujeitos normais e dos outros grupos psiquiátricos, o mesmo fazendo com os esquizofrênicos, os maníaco-depressivos, etc. Em *orientação vocacional*, o teste é construído em cima de grupos profissionais, dando o perfil desta profissão. Assim, o sujeito que vai ser testado responderá a este teste, que comporta séries de itens que caracterizam diferentes profissões, e será dito a ele que o seu perfil se aproxima de tal e tal grupo profissional e que, portanto, ele teria maiores chances de sucesso em tal profissão.

1.3 Avaliação Crítica

É neste tipo de testes onde impera o criticado empirismo cego, tipicamente predominando na área dos testes na época do domínio do behaviorismo, onde os testes eram criados e validados exclusivamente em termos da validade de critério. Mas há outros problemas graves com esta lógica de construção de testes. Vejamos:

A) Definição de grupos-critério

O ponto crucial dos testes referentes a critério consiste precisamente em selecionar grupos para servirem de critério para tomar decisões sobre que comportamentos ou itens definem este ou aquele grupo de sujeitos. Mas há aqui problemas, que se situam em dois níveis: na definição do próprio grupo critério e na medida deste grupo.

No que diz respeito à própria definição de grupo critério, temos, por exemplo, em psicologia clínica o caso das classificações psiquiátricas das doenças mentais. Existem ali as mais variadas classificações, desde as classificações médicas de Hipócrates (5º século antes de Cristo), as dos temperamentos de Galeno (2º século AD), as de Kraepelin (1883), em que se baseia por exemplo o MMPI, até as da

American Psychiatric Association de 1968 e das atuais. Se as definições de doenças mentais são diferentes, os testes construídos para avaliar as mesmas baseados em critério serão diferentes, porque os grupos critérios são diferentemente definidos e classificados. Quer dizer que um mesmo teste que inicialmente discriminava os grupos, depois já não descrimina mais porque as doenças foram classificadas diferentemente. Assim, o mesmo teste varia em diferentes ocasiões. O mesmo ocorre com o caso das profissões. Estas também se modificam com o passar do tempo, com a entrada de novas tecnologias, etc. Assim, os testes referentes a critério neste caso perdem o seu valor discriminativo porque as categorias profissionais mudaram de natureza.

Quanto à medida dos grupos critério, esta tem se mostrado demasiadamente precária. No caso do diagnóstico psiquiátrico, a base da medida é a opinião ou o diagnóstico subjetivo do psiquiatra. Há alguém mais competente para fazer tal diagnóstico do que o psiquiatra? Não. Mas qual é a confiança que se pode depositar em um diagnóstico subjetivo desta natureza? No caso das profissões, geralmente estas têm uma definição muito ampla demais para ser útil. Por exemplo, elas podem ser definidas como a de executivos, gerentes, assessores, etc. Mas existem executivos, gerentes, etc. de todo o tipo, dependendo da firma ser pública ou privada, grande ou pequena, de indústria ou de bancos, etc. Como parece difícil se poder definir a categoria profissional de uma maneira unívoca, fica impossibilitada a medida de quem pertence ou não à mesma. O problema não é se medir habilidades ou competências de um executivo, por exemplo, pois os testes referentes a critério não trabalham com tais conceitos. Trata-se de se poder classificar o sujeito numa categoria específica, o que fica complicado se esta não é claramente definida ou definível.

B) Falta de Significado Psicológico

Se fôssemos perguntar por que um teste referente a critério discrimina os grupos critério, a pergunta seria impertinente. Não interessa saber porque "cargas d'água" o teste discrimina, basta saber que ele o faz. E ele o faz porque foi construído para fazê-lo. Considere, por exemplo, o caso do MMPI. Este teste discrimina 10 categorias de sujeitos psiquiátricos. Agora, por que a escala de psicopatia discrimina os psicopatas de todos os outros? Analisando os itens desta escala individualmente, verificamos que eles falam de comportamentos os mais variados, desde preferências sexuais, sentimentos, situações de vida, dados socio-demográficos, etc. Então, uma série de conteúdos variados entram na definição da escala psiquiátrica. Assim, não é possível se decidir porque esta escala define o psicopata; apenas, se sabe que este sujeito responde a esta série heterogênea de itens de uma forma que é típica do psicopata. Assim, se quisesse construir uma forma paralela a esta escala, seria uma tarefa impossível, a não ser que se substituíssem os itens por sinônimos. Quer dizer, não há nestes testes nenhum fundamento de conteúdo psicológico que nos permita delimitar uma noção sobre o que é um psicopata, por exemplo.

Esta limitação dos testes referentes a critério é extremamente grave, porque os torna totalmente dependentes da situação em que foram construídos. Como, por exemplo, adaptar tais testes a outras culturas onde um sujeito psicopata pode ser assim por razões ou motivos muito diversos dos encontrados no grupo que serviu de

critério na construção original do teste? Ter-se-ia, praticamente, que reconstruir o teste na sua adaptação à nova situação. Disso resulta algo trágico, pois se as condições reais de vida dos sujeitos mudaram com o passar do tempo, de repente todo o banco de dados construído sobre profissões e grupos psiquiátricos se torna inutilizado, o que impediria o progresso do conhecimento. Tais testes, consequentemente, embora permitam simples diagnósticos para fins práticos, não permitem acumular o conhecimento e, portanto, a própria ciência psicológica.

C) Especificidade Situacional

Dada a falta de significado psicológico que os testes referentes a critério apresentam, eles se tornam totalmente dependentes dos grupos critérios em que se baseia a sua construção. A generabilidade destes testes se torna assim praticamente nula, tanto para a população da qual foram extraídos os grupos critério e mais ainda para populações diferentes, como é o caso do uso do teste assim construído em outras culturas. Por exemplo, em situação de seleção, as profissões mudam rapidamente com os avanços modernos da tecnologia (computador, Internet, etc.); assim; um teste referente a critério construído em cima de profissões que mudam de características se torna rapidamente obsoleto. Isto implica em que praticamente, para cada nova situação de testagem, um novo teste precisa ser construído e valendo somente para esta situação.

2. Testes Referentes a Construto

Os testes ou instrumentos psicológicos construídos com referência a construto partem da teoria psicológica e não de qualquer dado empírico. Eles pretendem representar ao nível dos comportamentos (itens) os traços latentes, os construtos, os conceitos psicológicos ou os processos psíquicos. Assim, o teste se constitui como uma hipótese empírica representando um traço latente, hipótese esta que deve ser demonstrada válida através da metodologia científica, isto é, o teste empírico.

Esta maneira de elaborar testes faz exigências cruciais sobre a teoria psicológica e esta afirmação já mostra as dificuldades que o pesquisador nesta área vai encontrar, dada a precariedade gritante desta teoria psicológica com referência a praticamente qualquer traço latente. A teoria psicológica, infelizmente, ainda labora no campo da fantasia mais do que em uma área onde se possam encontrar algumas balizas axiomatizadas. Assim, se torna difícil, diriam alguns impossível, utilizar a teoria, isto é, os traços latentes, como critério praticamente úteis para a construção dos instrumentos de medida psicológica. A bem da verdade, o ônus aqui cai sobre os ombros dos teóricos da Psicologia, que estão a dever uma teoria mais formalizada na área, embora o interesse neste campo da teoria esteja voltando com toda a força depois da desertificação que nele introduziu o behaviorismo ateórico. Com este retorno à teoria psicológica, os testes referentes a construto irão representar o grande campo da instrumentação psicológica. Infelizmente, o positivismo inculcado pelo behaviorismo em Psicologia ainda persiste renitente na maioria dos psicólogos e irá retardar um retorno mais rápido e avolumado da teoria em Psicologia. Por outro lado, o behaviorismo teve o mérito de afugentar da Psicologia o teorizar gratuito e

fantasioso característico do final do século passado e início deste e incentivar o teórico psicólogo a dar maior atenção aos dados empíricos, que necessariamente devem cimentar uma teoria científica.

De qualquer forma, a característica dos testes referentes a construto consiste no fato de que eles são construídos e validados para representar e medir traços latentes, isto é, conteúdos psicológicos, onde o significado das tarefas, itens, dita a inclusão ou não do item no teste, bem como dita a qualidade dele no teste, e não mais o simples fato de discriminar grupos diferentes de sujeitos. Este último aspecto, discriminação de grupos, será resultado necessário se os grupos de sujeitos diferem no traço que o teste mede. Aqui, pelo menos, se sabe, por que os grupos diferem e por que o teste os discrimina: porque os grupos diferem no traço latente. Além disso, sabendo-se que traço latente o teste mede, é possível se construírem testes paralelos que, com tarefas diferentes, meçam o mesmo traço latente, bem como é possível se adaptar um teste a culturas diferentes, desde que se saibam as características do traço latente na nova cultura. O ônus evidentemente cai todo sobre a definição de traço latente nas diferentes culturas. Enfim, o que dita a conduta para o teste é o traço latente, isto é, se o teste representa adequadamente este traço latente, e não o fato dele discriminar grupos critérios, embora este último deva ocorrer necessariamente se o teste mede corretamente o traço latente que difere nestes grupos.

Neste tipo de testes encontram-se a maioria dos testes psicológicos, os testes de inteligência e de aptidões, os inventários de personalidade e de psicopatologia e as escalas de atitude.

3. Testes Referentes a Conteúdo

A terminologia neste campo de testes referentes a conteúdo é bastante confusa. Infelizmente, Glaser (1963) introduziu nesta área uma expressão atualmente em voga, sobretudo na área da educação, que confundiu ainda mais o campo, chamando estes testes de "criterion-referenced testing". Infelizmente, porque tais testes não têm nada a ver com o que definimos acima sob o título de testes referentes a critério, como também não tem nada a ver com a validade de critério do modelo trinitário de Cronbach e Meehl (1955). Mas a expressão pegou e está sendo utilizada para caracterizar o tipo de testes que devem ser entendidos como testes referentes a conteúdo. Aliás, outras expressões são aqui utilizadas, embora estas sejam mais condizentes com o que estes testes pretendem, tais como testes referentes a domínio ou testes referentes a objetivos. Mayo (1945) inventou uma nova expressão, chamando estes testes de "mastery-referenced" (testes de mestria), porque se exige dos sujeitos a mestria de um conteúdo.

Testes referentes a conteúdo especificam um conteúdo e não tipos de pessoas, como fazem os testes referentes a critério e é neste contexto que deve ser entendida a diatribe entre normas referentes a critério e normas referentes a grupo, onde os escores do indivíduo são interpretados em função do escore do grupo obtido no mesmo teste. Em testes referentes a conteúdo, a interpretação do escore do sujeito se faz em função de um critério prévio e teoricamente estabelecido, como por exemplo, domínio de 80% de um conteúdo ensinado. Esta é também a razão de porque estes testes são chamados referentes a critério, a saber, porque existe um critério definido a priori de interpretação dos escores que os sujeitos recebem no teste. Aliás, os testes referentes a conteúdo são quase exclusivamente utilizados no âmbito da educação e aprendizagem (ensino, treinamento).

Este tipo de testes não visa discriminar sujeitos nem mesmo medir traços latentes, e sim verificar se os sujeitos atingem ou não um dado critério previamente definido, como o exemplo de 80% de domínio de um conteúdo de treinamento. A qualidade (a validade) destes testes vai depender exclusivamente deles serem amostras representativas de um conteúdo definido (de um domínio programático e de objetivos educacionais); por isso que as expressões testes referentes a conteúdo ou testes referentes a objetivos ou testes referentes a domínio são denominações mais apropriadas para estes testes, ao passo que as expressões testes referentes a critério ou testes referentes a mestria apenas acenam para características extrínsecas destas provas, especificamente elas se referem ao nível de realização dos resultados no teste.

4. Testes Comportamentais

A característica destes testes consiste no fato deles trabalharem exclusivamente com o comportamento observável e os estímulos ambientais (físicos, biológicos e sociais). O comportamento é entendido por Cunha (1975) como composto de estados orgânicos (aspectos funcionais como temperatura, cor, rigidez de um organismo), posturas (disposições espaciais estacionárias de partes do organismo umas em relação às outras) e movimentos (mudanças de posição espacial de partes do organismo em relação a outras).

4.1 Observação de Comportamento

Não se trata aqui propriamente de testes, pois estes constituem sempre um situação arranjada, enquanto a observação de comportamento é tipicamente feita em situação naturalística. Ela não visa avaliar traços latentes nem é seu interesse primário discriminar grupos de sujeitos. A observação de comportamento é de orientação etológica, isto é, ela está diretamente interessada em observar os comportamentos dos organismos com o objetivo de relatar sua ocorrência (freqüência), configuração, pretendendo eventualmente hipotetizar sobre sua origem filo ou ontogenética e fazendo comparações entre diferentes comportamentos do organismo e de diferentes organismos. Pode, inclusive, a partir dos comportamentos inferir para possíveis traços latentes como causas destes comportamentos. Contudo, a metodologia da observação de comportamento se debruça exclusivamente sobre o comportamento físico (verbal, motor) do ser vivo. Ela parte da definição de unidades comportamentais a serem observadas e registradas para chegar à elaboração de categorias de comportamento e da especificação das circunstâncias em que os comportamentos ocorreram. Como, no caso dos seres humanos, esta técnica não é utilizada gratuitamente, pelo simples prazer de observar, mas visa algum objetivo, a observação de comportamento pode ser considerada uma situação de teste (prova) na qual o observador está interessado em levantar categorias de comportamentos que o auxilie na avaliação de aspectos psicológicos do ser humano, como sua personalidade e habilidades.

4.2 Escalas Psicofísicas

As escalas psicofísicas têm como objetivo estabelecer uma relação de função entre estímulos ambientais (físicos, sociais) e o comportamento do indivíduo. De fato, elas visam escalonar magnitudes de estímulos com base nas reações (comportamentos) do organismo. A tarefa, então, consiste em definir a que nível de magnitude de dado estímulo tal ou tal resposta do organismo ocorre. Por exemplo, a que nível mínimo de decibéis o organismo começa a perceber o som ou a qual diferença mínima de decibéis o organismo reage diferencialmente.

5. Levantamentos: "Survey"

Novamente, não se trata propriamente de teste, embora testes de todo o tipo, inclusive observação do comportamento possam ocorrer numa situação de "survey". Trata-se mais de um delineamento de pesquisa na qual se deseja coletar informações as mais variadas sobre os sujeitos, tais como suas idéias, sentimentos, planos, crenças, origem social, educacional e financeira, etc. (Fink & Kosecoff, 1985). O levantamento de dados se constitui como uma situação arranjada, na medida em que os sujeitos são solicitados a responder a um questionário, que normalmente é o ponto central de tal técnica.

6. Novas Tecnologias

O avanço da informática está permitindo o desenvolvimento de tecnologias novas de coleta de dados em Psicologia, que se baseiam na própria capacidade do computador e na viabilidade de trabalhar com modelos multivariados. Vários capítulos deste livro exporão algumas destas tecnologias.

Bibliografia

Anastasi, A. (1988). *Psychological testing*. 6th ed. New York: Macmillan Publishing Company.

Campbell, D.T. (1957). A typology of tests, projective and otherwise. In D.N. Jackson & S. Messick (1967), *Problems in human assessment*. New York: McGraw-Hill Book Company, 190-194.

Cattell, R.B. & Warburton, F.W. (1967). *Objective personality and motivation tests*. Chicago, IL: University of Illinois Press.

Cronbach,L.J. & Meehl, P.E. (1955). Construct validity in psychological tests. In D.N. Jackson & S. Messick (1967), *Problems in human assessment*. New York: McGraw-Hill Book Company,57-77.

Cunha, W.H.A. (1975). O estudo etológico do comportamento animal. *Ciência e Cultura, 27*, 262-268.

Fink, A. & Kosecoff, J. (1985). *How to conduct surveys: A step-by-step guide*. Beverly Hills, CA: SAGE.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18*, 519-522.

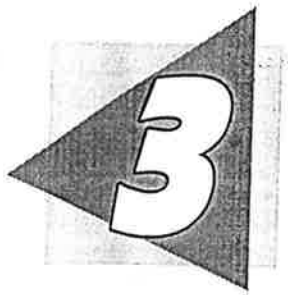
Kline, P. (1993). *The handbook of psychological testing*. London: Routledge.

Kraepelin, E. (1883). *Lehrbuch der Psychiatrie*.

Mayo, E. (1945). *The social problems of an industrial civilization*. Cambridge: Graduate School of Business Administration, Harvard University.

Rechtschaffen, A. & Mednick, S.A. (1955). The autokinetic word technique. *Journal of Abnormal Social Psychology, 51*, 386.

Sherriffs, A.C. (1948).The intuition questionnaire: A new projective test. *Journal of Abnormal Social Psychology, 43*, 326-337.



Testes Referentes a Construto: Teoria e Modelo de Construção

Luiz Pasquali

I. Introdução

Inicialmente, é importante alertar o leitor que a tecnologia aqui apresentada de elaboração de instrumentos psicológicos exige o conhecimento de algumas disciplinas ensinadas nas universidades, bagagem sem a qual dificilmente o pesquisador poderá se considerar apto a construir instrumentos psicológicos. Entre estas disciplinas salientam-se particularmente as seguintes, às quais este livro remete sem poder substituí-las, apenas indicando o momento no processo de elaboração do instrumento em que estas disciplinas têm seu espaço de aplicação:

- Psicometria: fundamental para a teoria da medida em Psicologia, particularmente, o conhecimento da Teoria da Resposta ao Item (TRI);
- disciplinas de teoria psicológica, tais como história e sistemas, teorias da personalidade, psicopatologia, psicologia social etc.; estas disciplinas são básicas para os procedimentos teóricos;
- disciplinas de delineamento de pesquisa científica; este conhecimento é fundamental para os procedimentos experimentais;
- disciplinas de estatística: estatística básica, análise de hipótese, análise fatorial; estes conhecimentos são decisivos nos procedimentos analíticos.

A teoria e o modelo de elaboração de instrumental psicológico apresentados neste capítulo são aplicáveis à construção de testes psicológicos de aptidão, de inventários de personalidade, de escalas psicométricas de atitude e do diferencial semântico. O modelo, que é detalhado na figura 3-1, se baseia em cima de três grandes pólos, que chamaremos de procedimentos teóricos, procedimentos empíricos (experimentais) e procedimentos analíticos (estatísticos).

O *pólo teórico* enfoca a questão da teoria que deve fundamentar qualquer empreendimento científico, no caso a explicitação da teoria sobre o construto ou objeto psicológico para o qual se quer desenvolver um instrumento de medida, bem como a operacionalização do construto em itens. Este pólo expõe a teoria do traço latente, bem como a explicitação dos tipos e categorias de comportamentos que constituem uma representação adequada do mesmo traço.

O *pólo empírico* ou experimental define as etapas e técnicas da aplicação do instrumento piloto e da coleta válida da informação para proceder à avaliação da qualidade psicométrica do instrumento.

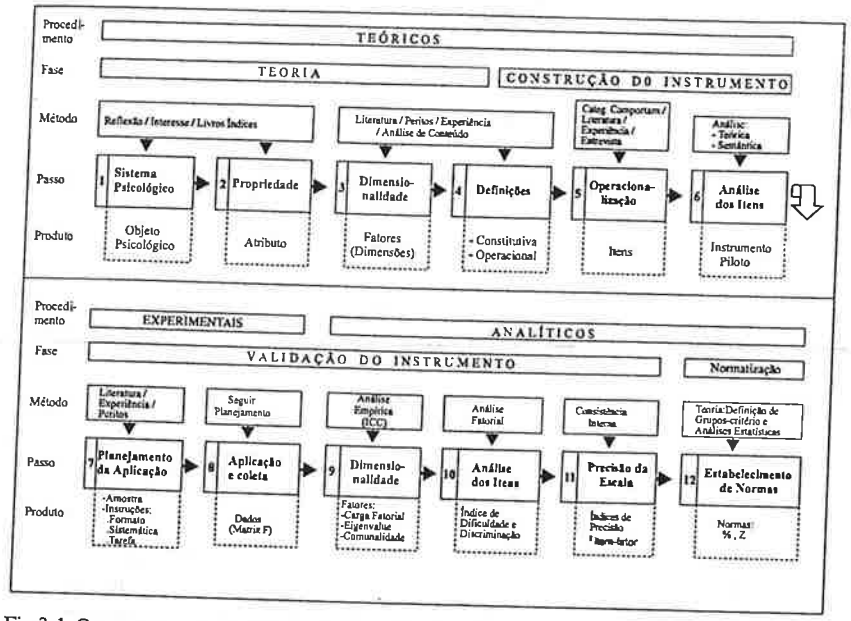


Fig.3-1. Organograma para elaboração de medida psicológica.

O *pólo analítico* estabelece os procedimentos de análises estatísticas a serem efetuadas sobre os dados para levar a um instrumento válido, preciso e, se for o caso, normatizado.

A figura 3-1 apresenta e detalha, para cada um destes três procedimentos, as etapas ou passos pelos quais se deve passar para se poder progredir sistematicamente na elaboração de um instrumento de medida psicológica baseado em construtos. Além disso, define, para cada passo, o método ou métodos a serem utilizados para superar o problema específico que constitui a tarefa a ser resolvida em cada passo, bem como o produto que decorre como resultado da solução do problema de cada passo. Além destes detalhes técnicos, a figura apresenta, para estes três procedimentos, uma meta-análise na qual se procura enquadrar e delimitar o evento ou eventos psicométricos que estão ocorrendo; tal fenômeno vem identificado sob a égide do rótulo "fase".

II. Procedimentos Teóricos

Os procedimentos teóricos devem ser elaborados para cada instrumento, dependendo, portanto, da literatura existente sobre o construto psicológico que o instrumento pretende medir. A teoria é, infelizmente ainda, a parte mais fraca da pesquisa e do conhecimento psicológicos, o que tem como conseqüência a precariedade dos atuais instrumentos psicométricos de medida nesta área. Na verdade, os instrumentos baseados numa teoria psicológica prévia mais elaborada (por exemplo,

"Edwards Personal Preference Schedule") não são dos melhores no mercado. Tal ocorrência explica, em parte, porque os psicometristas sistematicamente fogem da explicitação de uma teoria preliminar e iniciam a construção do instrumento pela coleta intuitiva e mais ou menos aleatória de uma amostra de itens, que dizem possuir "face validity", isto é, que parecem cobrir o traço para o qual eles querem elaborar o instrumento de medida. Embora isto não pareça muito científico, infelizmente é o que ocorre mais freqüentemente na construção de instrumental psicológico. A inexistência de teorias sólidas sobre um construto não deve ser desculpa para o psicometrista fugir de toda a especulação teórica sobre o mesmo. É obrigação dele levantar, pelo menos, toda a evidência empírica sobre o construto e procurar sistematizá-la e, assim, chegar a uma mini-teoria sobre o mesmo, a qual o possa guiar na elaboração de um instrumento de medida para o tal construto. Apesar do avanço e sofisticação estatísticos na Psicometria, parece ser esta fraqueza da base teórica que vem maculando a imagem dos procedimentos psicométricos na observação dos fenômenos psicológicos. Na verdade, com uma base teórica coerente e, quanto possível, completa, torna-se viável uma definição dos tipos e características dos comportamentos que irão constituir a representação empírica dos traços latentes e, assim, facilitar a tarefa do psicometrista em operacionalizá-los adequadamente (isto é, a construção dos itens se torna coerente e adequada).

De qualquer forma, a figura 3-2 detalha estes procedimentos teóricos.

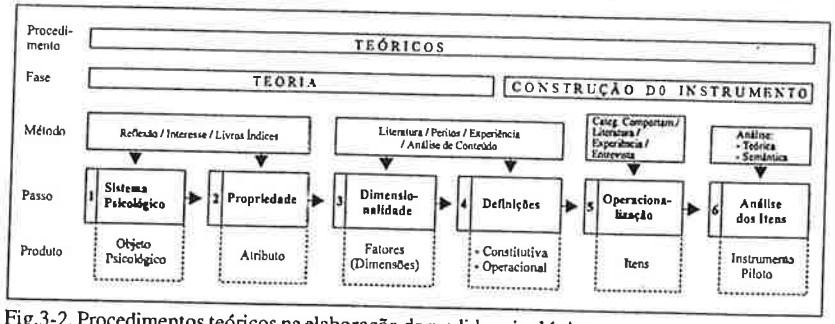


Fig.3-2. Procedimentos teóricos na elaboração da medida psicológica.

Nota Explicativa:

A terminologia em ciência e, diria, particularmente em Psicologia, infelizmente não é uniforme. Por isso, é útil conceituar preliminarmente certas expressões aqui utilizadas, como segue:

- *Sistema*: sinônimo de objeto, coisa, ser, entidade... que possui propriedades ou atributos. O sistema é definido não necessariamente pela natureza, mas pelo interesse do discurso e existente neste mundo do discurso.
- *Atributo*: propriedade, qualidade, aspecto, componente do objeto. Ele é caracterizado por ser mensurável num continuum de pontos de magnitude.
- *Magnitude*: qualidade de um sistema que pode assumir diferentes valores de quantidade, isto é, ela pode ser mais ou maior que (>) ou menor que (<).

- *Isomorfismo*: afirmação de correspondência entre propriedades do número (matemática) e quantidades das propriedades dos sistemas da natureza (física ou não).
- *Definição*: delimitar um conceito em termos de suas propriedades específicas. Ela é constitutiva ou formal se o conceito ou construto for definido em termos de outros construtos. Ela é operacional ou epistêmica se o conceito ou construto for definido em termos de fatos empíricos, da experiência ou observação.

2.1 O Sistema Psicológico

Qualquer sistema ou objeto que possa eventualmente ser expresso em termos observáveis é susceptível de se tornar um objeto para fins de mensuração. Acontece, porém, que um objeto em si não pode ser medido. Os objetos podem apenas ser enumerados. O que pode ser medido são as propriedades ou atributos de um objeto, desde que estes apresentem magnitudes, isto é, diferenças individuais, tais como intensidade, peso, altura, distância, etc. Por isso que estes atributos são geralmente chamados de variáveis, dado que não são invariantes entre sistemas individuais diferentes ou entre mesmos sistemas em diferentes ocasiões ou situações.

Se o sistema ou objeto representa o universo de interesse, o atributo dele constitui uma delimitação deste universo. O sistema realmente é definido pelo interesse do investigador. Como a ciência procura o conhecimento e não o poder ou a afirmação pessoal ou a política, então para o cientista não existe sistema privilegiado. Todo e qualquer sistema é digno e válido para ser conhecido. Obviamente, interesses políticos, sociais, pedagógicos, financeiros, etc. podem ditar a escolha de um objeto de estudo. Assim, a relevância de um sistema de estudo não é ditada pelo saber em si mas por fatores extrínsecos a ele; nem por isso estes fatores extrínsecos são negligenciáveis no contexto geral do universo da natureza e do homem, dado que o homem (pesquisador) está situado num contexto e tem suas prioridades em parte ditadas por este contexto. Assim, não há maior sabedoria em se estudar um grão de areia do que a sobrevivência do ser humano, embora para nós, seres humanos, esta última pareça bem mais relevante.

Enfim, o sistema representa o objeto de interesse, chamado também de objeto psicológico. A Psicometria enfoca como seu objeto específico as estruturas latentes, os traços psíquicos ou processos mentais, se quiser, que assim se constituem no seu objeto ou sistema direto de interesse. O sistema pode ser considerado de vários níveis, dependendo do interesse do pesquisador. Poder-se-ia falar de um sistema universal e de sistemas locais. O universal sendo a estrutura psicológica total do ser humano e os sistemas locais, os vários subsistemas de interesse. Assim, a inteligência poder ser considerada um subsistema dos processos cognitivos e estes da estrutura latente geral ou, mesmo, a inteligência digamos verbal pode ser considerada um sistema quando ela for o interesse imediato e na qual vários aspectos podem ser considerados, como a compreensão verbal e a fluência verbal. Sistema, portanto, constitui-se como tal enquanto objeto imediato de interesse dentro de um delineamento de estudo e não é uma entidade ontológica monolítica e unívoca.

Estes vários níveis de sistemas ocorrem mesmo nas coisas físicas. Assim, por exemplo, para o biólogo podem ser sistemas o organismo em sua totalidade ou partes dele, como é o sistema neurológico para o neurólogo, o sistema vascular para o cardiólogo, etc. O químico se interessa pelos elementos da tabela periódica, onde os seus sistemas naturais (água, ar, etc.) se reduzem a estes elementos de interesse deste profissional. O físico nuclear estuda seus sistemas reduzindo-os finalmente às partículas quark ("top", "bottom", "strange", etc.), às forças gluons (força forte, fraca, gravitacional) e aos processos leptons (elétrons, pósitrons). Em Psicologia, também encontramos tais níveis de sistemas. Considere, por exemplo, os processos cognitivos: Piaget e Spearman consideram a inteligência como uma grande estrutura (um sistema) que evolui geneticamente; os fatoristas consideram a inteligência ao nível de estruturas menores, quando falam de raciocínio verbal, numérico, abstrato, etc.; Sternberg vai ainda mais longe nesta elementarização dos sistemas (processos) cognitivos, buscando seus elementos no que ele chama de componentes cognitivos; e, finalmente, Newell e Simon levam ao extremo este elementarismo quando defendem os "elementary information process" (eip) como os elementos últimos dos processos cognitivos. A qualquer destes níveis, o pesquisador pode se colocar e definir este nível como o nível do sistema de seu interesse. Não é preciso ver oposições teóricas antagônicas nestes vários autores quanto aos processos cognitivos. Eles simplesmente se põem em horizontes diferentes e, por isso, vêem níveis diferentes de realidade, aliás, da mesma realidade. É apenas o exclusivismo, na verdade desnecessário, destes autores em afirmar que seu horizonte é o único ou o melhor para ver a realidade dos processos cognitivos. Isto vale, aliás, para qualquer outro processo psicológico, como a personalidade, por exemplo.

Enfim, o problema a ser resolvido neste passo praticamente se reduz a que o pesquisador, que pretende construir um instrumento, deve ter uma idéia, por mais vaga que seja, sobre o que é que ele quer trabalhar, para que tema da Psicologia ele está interessado em construir um instrumento de medida e pesquisa. Este problema é, evidentemente, mais aparente em aluno de pós-graduação, ao qual se apresenta a necessidade de elaborar uma dissertação no final do curso e ainda não tem idéia sobre que assunto em Psicologia ele quer desenvolver esta tese. Na falta de qualquer outra indicação ou interesse específico, tal indivíduo pode se dirigir aos livros índices onde estão elencados os principais trabalhos que se vêm fazendo em Psicologia. Para o psicólogo há uma série de tais livros, sendo o mais útil o *Psychological Abstracts*, que é publicado mensalmente e onde aparece a quase totalidade dos artigos e trabalhos feitos em Psicologia a nível mundial. Além deste, há o *Educational Index*, para os interessados na psicologia aplicada à educação, *o Index Medicus*, para os interessados em psicologia clínica, e o *Sociological Index*, onde aparecem temas referentes à psicologia social. Na falta de tais fontes ou se ainda assim, o tema não surgiu na percepção do pesquisador (aluno), este pode recorrer a peritos, que, no caso do aluno, é normalmente seu orientador. Enfim, o problema a ser resolvido neste passo consiste em se ter uma idéia, um tema, um assunto para pesquisar. A este tema chamamos de objeto psicológico, que representa o produto esperado deste passo na elaboração do instrumento. Agora, este sistema escolhido pode ser mais amplo ou mais restrito, como vimos acima ao falarmos dos diferentes níveis de sistemas. Obviamente, quanto mais restrito ou elementar for o sistema, mais fácil se torna a

construção de um instrumento de medida. Por isso, é relevante definir como sistema psicológico um processo ou traço latente o mais próximo do interesse direto do pesquisador; o que tipicamente significa em definir um sistema o mais elementar possível, dentro do interesse. Sistemas vagos e gerais dificultam depois sua operacionalização para fins de pesquisa empírica, como é a construção de um instrumento de medida.

2.2 A Propriedade do Sistema Psicológico

O sistema, já dissemos, não constitui objeto direto de mensuração, mas sim suas propriedades ou atributos que são os vários aspectos que o caracterizam. Por exemplo, o sistema físico se apresenta com os atributos de massa, comprimento, etc. Similarmente, a Psicometria concebe os seus sistemas como possuidores de propriedades/atributos que definem os mesmos, sendo estes atributos o foco imediato de observação/medida. Assim, a estrutura psicológica apresenta atributos do tipo processos cognitivos, processos emotivos, processos motores, etc. A inteligência, como subsistema, pode apresentar atributos de tipo raciocínio verbal, raciocínio numérico, etc. O sistema se constitui como objeto hipotético que é abordado (conhecido) através da pesquisa de seus atributos.

O problema específico deste passo consiste em se passar de um objeto psicológico, normalmente amplo demais para pesquisar, para a delimitação do ou dos aspectos específicos dele os quais se deseja estudar e para os quais se quer construir um instrumento de medida. De fato, qualquer sistema apresenta ilimitado número de propriedades. A rosa, por exemplo, tem perfume, cor, peso, tamanho, beleza, ritmo de crescimento, etc. É relevante, para se poder escolher ou construir um instrumento de medida, definir qual ou quais propriedades do sistema serão objeto de estudo. Por exemplo, se meu interesse se focaliza sobre a criança. Não é possível estudar, de uma só vez, tudo sobre a criança. Então, tenho que me decidir por um aspecto mais restrito referente à criança, o qual enfim vou pesquisar. Assim, da criança posso estudar o seu desenvolvimento psicomotor, o desenvolvimento cognitivo, o desenvolvimento da linguagem, a enurese, a timidez, a agressividade, etc. Qual destes ou outros aspectos estou presentemente e diretamente mais interessado? Pois, este ou estes aspectos constituem a propriedade do objeto criança que presentemente quero abordar. A estes aspectos escolhidos chamamos de atributo. Ou, o seu interesse pode se focalizar sobre a inteligência. Esta de fato já é em si mesma uma propriedade do sistema homem. Mas ela pode ser igualmente considerada um subsistema complexo, apresentando várias propriedades específicas dela, tais como, raciocínio verbal, raciocínio numérico, raciocínio abstrato, memória, percepção espacial, etc. Para se definir um instrumento de medida, é preciso se decidir qual ou quais destas propriedades serão o objeto imediato de interesse. Ademais, à medida que o conhecimento sobre o sistema cresce, cresce também o número de novas propriedades descobertas, que, por sua vez, podem se tornar novos subsistemas de interesse, vez que se vão descobrindo propriedades diversas dentro destes novos subsistemas. Enfim, é de importância para se prosseguir sem transtornos e desvios de rumo que se defina claramente e preliminarmente a ou as propriedades do sistema de interesse que se quer estudar. Tal definição evita que se misture, no prosseguimento do processo, alhos e bugalhos,

como, por exemplo, utilizar uma amostra de itens que mais medem aspectos de conhecimento de vocabulário, quando, de fato, se queria atingir o raciocínio verbal. Com isso, também não se está afirmando que entre tais propriedades de um sistema não haja correlações. Antes, pelo contrário, relações e interações entre as propriedades de um mesmo sistema são uma suposição não somente legítima mas provável. Contudo, o que se está afirmando é que é preciso partir com conceituações e definições claras e precisas, bem como delimitadas, dado que a capacidade de conhecimento humano não é abrangente.

Como se decidir por este ou aquele aspecto? Novamente, recorro ao meu interesse, à ajuda dos livros índices e aos peritos (concretamente, ao meu orientador, se for aluno).

2.3 Dimensionalidade do Atributo

Se os dois primeiros passos acima descritos possam parecer, para muitos, um mero exercício acadêmico, o terceiro passo e os demais a seguir já não são tão simples, pois os problemas que eles apresentam já são bem mais complexos.

A dimensionalidade do atributo diz respeito à sua estrutura interna, semântica. O atributo constitui uma unidade semântica única ou é ele uma síntese de componentes distintos ou até independentes? Deve ele ser concebido como uma dimensão homogênea ou deve-se nele distinguir aspectos diferenciados? A resposta a este problema obviamente deve vir ou da teoria sobre o construto e/ou dos dados empíricos disponíveis sobre ele, sobretudo dados de pesquisas que utilizaram a análise fatorial na análise dos dados, pois o que está em jogo aqui é a questão de decidir se o construto é uni ou multifatorial. Os fatores que compõem o construto (o atributo) são o produto deste passo. Por exemplo: Tenho como objeto psicológico os processos cognitivos; a propriedade deste objeto psicológico que estou interessado em estudar é a inteligência verbal. Pergunta-se: é esta inteligência verbal um construto único ou devo distinguir nele componentes diferentes? Os dados empírico disponíveis me mostram que a inteligência verbal é composta por, pelo menos, dois fatores bem distintos e praticamente independentes, a saber, compreensão verbal e fluência verbal. Consequentemente, se quiser pesquisar a inteligência verbal e construir para tal um instrumento de medida, não poderei prescindir de conhecer e levar em conta este fato de que esta inteligência apresenta dois fatores distintos, cuja medida de ambos exige instrumentos diferentes. Claro, posso me decidir por estudar somente a inteligência verbal compreendida sob seu aspecto de compreensão verbal e prescindir de me preocupar com a fluência verbal. Tudo bem, mas neste caso o meu atributo de interesse de estudo não é mais a inteligência verbal e sim a compreensão verbal. Mesmo tomando esta decisão de somente querer estudar a compreensão verbal, não fico escusado de expor a teoria sobre a inteligência verbal em sua totalidade e, em seguida, justificar minha decisão pelo estudo de apenas um aspecto dela. Evidentemente, esta justificativa pode ser e será suficiente o meu interesse específico por tal aspecto da inteligência. Isto é, eu devo saber o que estou fazendo, e demonstro isso na exposição da teoria que faço sobre o construto inteligência verbal.

Nestes dois passos, propriedade e dimensionalidade, entramos no ponto mais crítico na caminhada para a elaboração dos instrumentos psicológicos, porque toda esta parte resulta essencialmente da teoria psicológica, a qual concebe, define e estrutura

os construtos psicológicos. A tarefa da construção da teoria psicológica não é tarefa específica do psicometrista e sim do psicólogo teórico. O psicometrista deveria poder contar com esta teoria e com base nela fundamentar a construção dos instrumentos de medida. A existência de teorias ou fantasias as mais variadas sobre praticamente qualquer construto em Psicologia torna a tarefa do psicometrista quase uma tragédia quando quer construir instrumentos para medir construtos sobre os quais os psicólogos não se entendem. Desta sorte, o psicometrista acaba se decidindo em construir um instrumento para medir um construto concebido *segundo* algum psicólogo. E ali você tem uma fauna enorme de psicólogos teóricos, desde os behaviorista até os dialéticos, que falam linguagens quase totalmente estranhas um em relação ao outro. Infelizmente esta é a situação da teoria psicológica atual. Para caricaturar, imagine o seguinte: um físico vai construir um instrumento para medir o comprimento de objetos físicos. Mas, se para poder efetuar tal empreendimento, ele tivesse que decidir sobre "bem, comprimento entendido segundo quem?" Tal pergunta careceria de sentido e seria ridícula fosse ela feita sobre comprimento ou outras propriedades da matéria (pelo menos, na sua grande maioria). Mas, na caso do psicometrista, tal pergunta infelizmente é corriqueira, qualquer que seja o construto que ele queira estudar e medir, o que vem a mostrar o estado primitivo em que vive a teoria psicológica. Esta precariedade da teoria psicológica é a principal responsável pela fuga, por parte dos psicometristas, de basear a construção dos instrumentos psicológicos numa teoria prévia e testá-los em seguida através da metodologia científica. Esta fuga acarreta que o psicometrista parte de uma coleção atabalhoada de itens para em seguida ver o que eles estão medindo, se alguma coisa psicológica relevante.

Este estado de coisas deveria e deve obrigar o psicometrista a expor ou elaborar uma mini-teoria sobre o que ele entende pelo construto que pretende medir. Felizmente, já existe razoável abundância de dados empíricos sobre muitos construtos psicológicos, com base nos quais o psicometrista poderá desenvolver uma tal mini-teoria do construto, a qual irá guiar a construção do seu instrumento de medida. Os dados empíricos que serão coletados através do instrumento assim construído irá decidir se sua mini-teoria tem ou não alguma consistência. Isso não é uma tragédia, é a própria lógica da pesquisa empírica, isto é, a testagem empírica que pode ou não confirmar a validade de uma teoria: a verdade científica é sempre relativa, nunca será um dogma, e portanto sempre reformável.

2.4 Definição dos Construtos

Decididas a propriedade e suas dimensões, é preciso conceituar detalhadamente estes construtos, novamente baseando-se na literatura pertinente, nos peritos da área e na própria experiência. O problema deste passo é, portanto, a conceituação clara e precisa dos fatores para os quais se quer construir o instrumento de medida. A tarefa aqui é dupla, tendo como resultado dois produtos, a saber, as definições constitutivas e as definições operacionais dos construtos.

2.4.1 Definição Constitutiva

Um construto definido através de outros construtos representa uma definição constitutiva. Neste caso, o construto é concebido em termos de conceitos

próprios da teoria em que ele se insere. Definição constitutiva é a que tipicamente aparece como definição de termos em dicionários e enciclopédias: os conceitos são ali definidos em termos de outros conceitos; isto é, os conceitos, que são realidades abstratas, são definidos em termos de realidades abstratas. Por exemplo, se defino inteligência verbal como a capacidade de compreender a linguagem, estou diante de uma definição constitutiva, porque capacidade de compreender constitui uma realidade abstrata, um construto, um conceito.

As definições constitutivas são de extrema importância no contexto da construção dos instrumentos de medida, porque elas situam o construto exata e precisamente dentro da teoria deste construto, dando, portanto, as balizas e os limites que ele possui. Enfim, estas definições caracterizam o construto, dando as dimensões que ele deve assumir no espaço semântico da teoria em que está incluído. Assim, se defino assertividade como a capacidade de dizer não, a capacidade de expressar livremente sentimentos positivos e negativos, a capacidade de expor idéias sem receio, etc., estou dando os limites semânticos que este conceito deve respeitar dentro da minha teoria de assertividade. Definições desta natureza põem limitações definidas sobre o que devo explorar quando for medir este construto, limitações não somente em termos de fronteiras que não podem ser ultrapassadas, mas mais ainda em termos de fronteiras que devem ser atingidas. De fato, normalmente um instrumento que mede um construto não chega a cobrir toda amplitude semântica de um conceito. Assim, boas definições constitutivas vão me permitir em seguida avaliar a qualidade do instrumento que mede o construto em termos de quanto desta extensão semântica do mesmo é coberta pelo instrumento, surgindo daí instrumentos melhores e piores na medida em que medem mais ou medem menos da extensão conceptual do construto, extensão esta delimitada pela definição constitutiva deste mesmo construto.

2.4.2 Definição Operacional

Com as definições constitutivas nós estamos ainda no terreno da teoria, do abstrato. Um instrumento de medida já é uma operação concreta, empírica. A passagem do terreno abstrato para o concreto é precisamente viabilizada pelas definições operacionais dos construtos. Este é, talvez, o momento mais crítico na construção de medidas psicológicas, pois é aqui que se fundamenta a validade destes instrumentos; é aqui que se baseia a legitimidade da representação empírica, comportamental, dos traços latentes (os construtos). Duas preocupações são relevantes e decisivas neste momento: 1) as definições operacionais dos construtos devem ser realmente operacionais e 2) elas devem ser o mais abrangentes possível dos construtos.

Primeiramente, as definições operacionais devem ser realmente *operacionais*. Esta tautologia é proposital, porque se peca demais neste particular. Uma definição de um construto é operacional quando o mesmo construto é definido, não mais em termos de outros construtos, mas em termos de operações concretas, isto é, de comportamentos físicos através dos quais o tal construto se expressa. Assim, se defino inteligência verbal como a capacidade de compreender uma palavra ou, mesmo, compreender uma frase, estou diante de uma definição constitutiva e não operacional. Isto porque compreender não é um comportamento, mas um construto. Seria uma definição operacional de compreensão da frase, *reproduzir a frase* com

outras palavras. Mager (1981) dá uma fórmula simples e perfeita para decidir se a definição é ou não operacional. Ela é operacional se você puder dizer ao sujeito : "*vá e faça ...*". Assim, se defino inteligência verbal como compreender uma frase, o que devo pedir ao sujeito para fazer, uma vez que "vá e compreenda..." não diz ao sujeito nada que ele possa fazer? Ao passo que dizer "vá e reproduza a frase" indica claramente o que o sujeito deve fazer, como deve se comportar, e, portanto, esta última é uma definição operacional, pois ela define comportamentos que devem ocorrer, enquanto compreender a frase não indica nenhum comportamento concreto específico a ser exibido por parte do sujeito.

Em segundo lugar, a definição operacional deve ser o mais *abrangente* possível do construto. Nenhuma definição operacional esgota a amplitude semântica de um construto; assim, podem haver definições operacionais mais o menos abrangentes do mesmo construto e esta grandeza de abrangência, evidentemente, fala da boa, má ou pior qualidade da definição operacional, o que vai obviamente repercutir sobre o instrumento de medida do construto que será baseado nesta definição operacional do mesmo construto. Aliás, uma definição operacional pode ser perfeitamente operacional e também perfeitamente equivocada ou errada, quando esta não cobrir nada do espaço semântico próprio do construto. Assim definir inteligência verbal como desenhar círculos na areia constitui uma definição perfeitamente operacional, pois todo o mundo entende quando se manda desenhar círculos na areia; contudo, apesar de operacional, ela é uma definição perfeitamente equivocada de inteligência verbal, pois este comportamento de desenhar círculos na areia não tem nada a ver com o construto em questão. Disto segue que as definições operacionais podem representar um construto numa escala que expressa uma proporção de coincidência entre construto e definição operacional que vai de 0 a 1; sendo 0, quando a definição não cobre nada do construto e 1, quando ela cobrir 100% do espaço semântico do construto. Como já dissemos, cobrir 100% do construto nenhuma definição operacional será capaz, mas quanto maior covariância existir entre construto e definição operacional, maior qualidade se deve atribuir a esta definição do construto e, por conseqüência, maior chance terá o instrumento, que de tal definição resulta, de ser superior em qualidade. Dizemos maior chance, porque a qualidade do instrumento não depende unicamente de boas definições operacionais, embora sem a boa qualidade destas o instrumento já começa de saída a ser inferior. A seguinte ilustração deixa visualizar esta problemática da qualidade de representação comportamental de diferentes definições operacionais do construto compreensão verbal:

	Compreensão verbal
Extensão semântica total	
Definição Operacional	
Dizer que compreendeu	
Desenhar círculos	
Escrever a frase	
Reproduzir a frase	
...	...

Para garantir melhor cobertura do construto, as definições operacionais deverão especificar e elencar aquelas categorias de comportamentos que seriam a representação comportamental do construto. Quanto melhor e mais completa for esta especificação, melhor será a garantia de que o instrumento que resultar para a medida do construto será válido e útil. Por exemplo, quais seriam as categorias de comportamentos que expressariam comportamentalmente a compreensão verbal? Seriam tais como: reproduzir texto, dar sinônimos e antônimos, explicar o texto, sublinhar alternativas, etc. Quanto mais completa esta listagem de categorias comportamentais, mais próximo estou da construção do instrumento, porque o próximo passo será simplesmente expressar estas categorias em tarefas unitárias e específicas (os itens) e o instrumento piloto está construído. Por isso, nunca é demais gastar tempo na implementação detalhada das definições operacionais do construto.

Onde vou me inspirar para realizar adequadamente esta tarefa? Novamente, os métodos a serem utilizados para resolver o problema deste passo de construção de medidas psicológicas são a literatura pertinente sobre o construto, a opinião de peritos na área, a experiência do próprio pesquisador, bem como a análise de conteúdo do construto. Torna-se aqui, como se vê, indispensável o conhecimento aprofundado da literatura sobre o construto, bem como das técnicas de análise de conteúdo.

É bom lembrar neste contexto de que os instrumentos de medida psicológica visam medir traços latentes. Mas como medir traços latentes que são impérvios à observação empírica que é o método da ciência? Estamos aqui nos defrontando com o problema da representação: qual é a maneira adequada de se representar estes atributos latentes para que possam ser cientificamente abordados? Embora o problema pareça, e é na verdade, grave, ele não é específico da Psicometria, ele ocorre na própria física com a teoria quântica, por exemplo. Como o comportamento representa estes traços latentes? É precisamente o problema que as definições operacionais precisam resolver.

2.5 Operacionalização do Construto

Este é o passo da construção dos itens, que são a expressão da representação comportamental do construto, a saber, as tarefas que os sujeitos terão de executar para que se possa avaliar a magnitude de presença do construto (atributo).

2.5.1 Fontes dos Itens

Se os passos até aqui discutidos foram adequadamente resolvidos, nós estamos agora diante das categorias comportamentais que expressam o construto de interesse, as quais dão praticamente a resposta à construção dos itens. Além disso, podemos apelar para outras duas fontes de itens: a entrevista e outros testes que medem o mesmo construto. A entrevista consiste em pedir a sujeitos representantes da população para a qual se deseja construir o instrumento para opinarem em que tipo de comportamentos tal construto se manifesta. Por exemplo, se meu desejo é construir um instrumento sobre assertividade, posso me dirigir a representantes da população e perguntar: "como é para você uma pessoa assertiva"? De uma pesquisa desta

natureza pode surgir uma grande riqueza de comportamentos que expressam assertividade e que podem ser aproveitados como itens do instrumento. Ademais, posso me inspirar em itens que compõem outros instrumentos disponíveis no mercado e que medem o mesmo construto no qual estou interessado. Assim, temos três fontes preciosas para a construção dos itens:

- literatura: outros testes que medem o construto;
- entrevista: levantamento junto à população alvo;
- categorias comportamentais: definidas no passo das definições operacionais.

É importante notar que no processo de elaboração do instrumento como o temos exposto, os itens não são mais coletados a esmo ou chutados, mas eles são elaborados ou, pelo menos, selecionados em função das definições operacionais de um construto que foi exaustivamente analisado em seus fundamentos teóricos e nas evidências (dados) empíricas disponíveis. Então, não é qualquer item que pareça medir o construto que é aceito, mas somente aquele que corresponde às definições teóricas (constitutivas) e às definições operacionais do mesmo. Não é mais a malfadada "face validity" que impera na seleção dos itens e sim a sua pertinência (a esta altura, obviamente, ainda teórica) ao contexto teórico do construto. Aliás, os itens não são selecionados ou pescados, eles são é construídos para representar comportamentalmente o construto de interesse.

2.5.2 Regras para Construção de Itens

Dadas as fontes que baseiam a construção dos itens, é preciso dar agora algumas regras ou critérios fundamentais para a elaboração adequada dos próprios itens. Estas regras se aplicam, em parte, à construção de cada item individualmente, e em parte ao conjunto dos itens que medem um mesmo construto. Ademais, dependendo do tipo de traço a ser medido seja de aptidão ou de personalidade, algumas das regras se aplicam e outras não.

a) Critérios para a construção dos itens:

1) *Critério comportamental*: o item deve expressar um comportamento, não uma abstração ou construto. Segundo Mager (1981), o item deve poder permitir ao sujeito uma ação clara e precisa, de sorte que se possa dizer a ele "vá e faça". Assim 'reproduzir um texto' é um item comportamental ('vá e reproduza...'), ao passo que 'compreender um texto' não o é, pois o sujeito não sabe o que fazer com 'vá e compreenda...'.

2) *Critério de objetividade ou de desejabilidade ou preferência*: para o caso de escalas de aptidão, os itens devem cobrir comportamentos de fato, permitindo uma resposta certa ou errada. O sujeito respondente deve poder mostrar se conhece a resposta ou se é capaz de executar a tarefa proposta. Assim, por exemplo, se você quer saber se o sujeito entende o que seja "abstêmio", faz mais sentido pedir a ele que dê um sinônimo do que pedir que diga se entendeu ou não. Ao contrário, para o caso de personalidade e das atitudes em geral, os itens devem cobrir comportamentos desejáveis (atitude) ou característicos (personalidade). O respondente, neste caso, deve poder concordar ou discordar ou opinar sobre se tal comportamento convém ou não para ele, isto é, os itens devem expressar desejabilidade ou preferência. Não

existem neste caso respostas certas ou erradas; existem sim diferentes gostos, preferências, sentimentos e modos de ser.

3) *Critério da simplicidade*: um item deve expressar uma única idéia. Itens que introduzem explicações de termos ou oferecem razões ou justificativas são normalmente confusos porque introduzem idéias variadas e confundem o respondente. Por exemplo: "Gosto de feijão porque é saudável". O sujeito pode de fato gostar de feijão mas não porque seja saudável; assim, ele não saberia como reagir a tal item: se porque o feijão é gostoso ou porque é saudável. O item exprime duas idéias. O mesmo vale para "a maçã é gostosa e saudável".

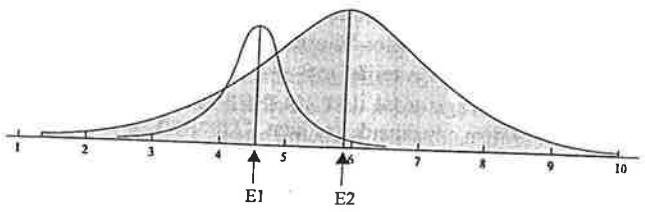
4) *Critério da clareza*: o item deve ser inteligível até para o estrato mais baixo da população alvo; daí, utilizar frases curtas, com expressões simples e inequívocas. Frases longas e negativas incorrem facilmente na falta de clareza. Com referência às frases negativas: normalmente elas são mais confusas que as positivas; consequentemente, é melhor afirmar a negatividade do que negar uma afirmação. Por exemplo: fica mais inteligível dizer "detesto ser interrompido" do que "não gosto de ser interrompido" ou em vez de "não me sinto feliz" é melhor dizer "sinto-me infeliz". Neste contexto, é preciso igualmente fazer atenção em não utilizar gírias, porque estas não são normalmente inteligíveis para todos os membros de uma população alvo do instrumento, além de tipicamente ofender o estrato mais sofisticado da mesma população, o que pecaria contra o critério numero 10. Contudo, o linguajar típico da população alvo deve ser utilizado na formulação dos itens; assim, são admissíveis e são mais apropriadas expressões conhecidas por tal população, ainda que elas possam parecer lingüisticamente menos castiças. A preocupação aqui é a compreensão das frases (que representam tarefas a serem entendidas e se possível resolvidas), não sua elegância artística.

5) *Critério da relevância* (pertinência, saturação, unidimensionalidade, correspondência): a expressão (frase) deve ser consistente com o traço (atributo, fator, propriedade psicológica) definido e com as outras frases que cobrem o mesmo atributo. Isto é, o item não deve insinuar atributo diferente do definido. O critério diz respeito à saturação que o item tem com o construto, representada pela carga fatorial na análise fatorial e que constitui a covariância (correlação) entre o item e o fator (traço). Veja o seguinte exemplo: seja o construto "compreensão verbal" definido como compreender o significado de palavras e frases. Dos três itens abaixo, um pertinente, outro mais ou menos e um é impertinente:

– Reproduzir a frase com as próprias palavras	→	pertinente
– Decorar uma sentença	→	pouco pertinente
– Falar em voz alta	→	impertinente.

6) *Critério da precisão*: o item deve possuir uma posição definida no contínuo do atributo e ser distinto dos demais itens que cobrem o mesmo contínuo. Este critério supõe que o item possa ser localizado numa escala de estímulos; em termos de Thurstone, diríamos que o item deve ter uma posição escalar modal definida e um desvio padrão reduzido. Em termos da TRI, este critério é representado pelos parâmetros 'b' (dificuldade) e "a" (discrimina-

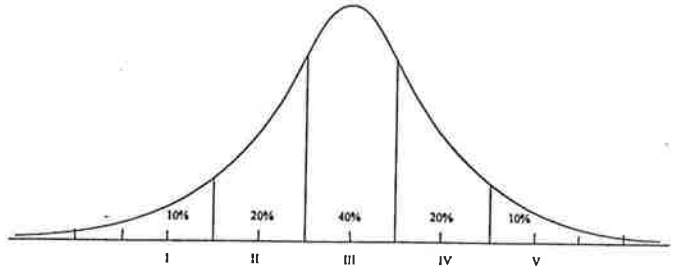
ção) e pode realmente ser avaliado definitivamente somente após coleta de dados empíricos sobre os itens. Por exemplo, na escala de Thurstone abaixo, o item E1 é muito preciso, enquanto o E2 é impreciso.



7) *Critério da variedade*: dois aspectos especificam este critério:
   (1) variar a linguagem: uso dos mesmos termos em todos os itens confunde as frases e dificulta diferenciá-las, além de provocar monotonia, cansaço e aborrecimento. Exemplo: o EPPS (Edwards Personal Preference Schedule) em inglês começa quase todas as suas 500 frases com a expressão "I like...". Depois de tantos "I like", qualquer sujeito deve se sentir saturado!
   (2) no caso de escalas de preferências: formular a metade dos itens em termos favoráveis e metade em termos desfavoráveis, para evitar erro da resposta estereotipada à esquerda ou à direita da escala de resposta. É a recomendação que Likert já dava em 1932.
8) *Critério da modalidade*: formular frases com expressões de reação modal, isto é, não utilizar expressões extremadas, como 'excelente', 'miserável', etc. Assim, ninguém é *infinitamente* inteligente, mas a maioria é *bastante* inteligente. A intensidade da reação do sujeito é dada na escala de resposta. Se o próprio item já vem apresentado em forma extremada, a resposta na escala de respostas já está viciada. Assim, se pergunto ao sujeito se está pouco ou muito de acordo (numa escala, por exemplo, de 7 pontos que vai de desacordo total a acordo total), um item formulado extremado tal como "meus pais são a melhor coisa do mundo" dificilmente receberia resposta 7 (totalmente de acordo) por parte da maioria dos sujeitos da população alvo, simplesmente porque a formulação é exagerada. Se em lugar dela eu usasse uma expressão mais modal, tal como "eu gosto dos meus pais", as chances de respostas mais variadas e inclusive extremadas (resposta 7) seriam de se esperar.
9) *Critério da tipicidade*: formar frases com expressões condizentes (típicas, próprias, inerentes) com o atributo. Assim, a beleza não é pesada, nem grossa, nem nojenta.
10) *Critério da credibilidade* (face validity): O item deve ser formulado de modo a que não apareça sendo ridículo, despropositado ou infantil. Itens com esta última caracterização fazem o adulto se sentir ofendido, irritado ou coisa similar. Enfim, a formulação do item pode contribuir e contribui (Nevo, 1985; Nevo & Sfez, 1985) para uma atitude desfavorável para com o teste e assim aumentar os erros (vieses) de resposta. Este tema, às vezes, é discutido sob o

que se chama de validade aparente (*face validity*), que não tem nada a ver com a validade objetiva do teste, mas pode afetar negativamente a resposta ao teste, ao afetar o indivíduo respondente.

b) Critérios referentes ao conjunto dos itens (o instrumento todo):
   11) *Critério da amplitude*: este critério afirma que o conjunto dos itens referentes ao mesmo atributo deve cobrir toda a extensão de magnitude do continuum deste atributo. Critério novamente satisfeito pela análise da distribuição dos parâmetros 'b' da TRI. A razão disto é que um instrumento deve poder discriminar entre sujeitos de diferentes níveis de magnitude do traço latente, inclusive entre os que possuem um traço alto quanto entre os que possuem um traço pequeno, e não somente entre os de traço alto e traço baixo.
   12) *Critério do equilíbrio*: os itens do mesmo contínuo devem cobrir igualmente ou proporcionalmente todos os segmentos (setores) do contínuo, devendo haver, portanto, itens fáceis, difíceis e médios (para aptidões) ou fracos, moderados e extremos (no caso das atitudes). De fato, os itens devem se distribuir sobre o contínuo numa distribuição que se assemelha à da curva normal: maior parte dos itens de dificuldade mediana e diminuindo progressivamente em direção às caudas (itens fáceis e itens difíceis em número menor). A razão deste critério se encontra no fato de que a grande maioria dos traços latentes se distribuem entre a população mais ou menos dentro da curva normal, isto é, a maioria dos sujeitos possuem magnitudes medianas dos traços latentes, sendo que uns poucos possuem magnitudes grandes e outros magnitudes pequenas. Assim, a distribuição dos itens num instrumento deve ser mais ou menos segundo a curva normal, como mostrado na figura abaixo, onde se diz que 10% dos itens devem ter dificuldade mínima ou máxima, 40% dificuldade mediana, etc.:



Distribuição percentual dos intens em 5 faixas de dificuldade

2.5.3 Quantidade de Itens

Para se cobrir a totalidade ou a maior parte ou, pelo menos, grande parte da extensão semântica do construto, explicitada nas definições constitutivas, normalmente se exige, no instrumento final, um número razoável de itens. Que é um número razoável? O bom senso de quem trabalha nesta área sugere que um construto, para ser bem representado, necessita de cerca de 20 itens. Há, evidentemente, construtos muito simples que dificilmente necessitam de tal número itens, sendo sufici-

entes apenas uma meia dúzia ou menos deles. Por exemplo, satisfação com o salário. Quantas maneiras há de se verificar tal satisfação? Parece exagerado perguntar 20 vezes ao sujeito se está satisfeito com o seu salário. Posso, sim, perguntar se ele está contente com a quantia, com o poder de compra, com a pontualidade de entrega, e alguns mais aspectos. Mas parece difícil descobrir umas 20 maneiras de estar satisfeito com o salário. Entretanto, em sua grande maioria, os traços latentes possuem uma gama bem maior de aspectos e, por isso, exigem maior número de itens para ser adequadamente representados.

Se o número final de itens, isto é, depois que o instrumento passou por todas as fases de construção e validação, deve ser em torno de 20, pergunta-se com quantos itens é preciso começar para que no final possamos salvar 20? A resposta dada ao contexto da psicometria tradicional positivista é a de que se deve começar com, pelo menos, o triplo de itens para se poder assegurar, no final, um terço deles. Esta resposta se deve ao modo positivista ou ateórico de construir instrumentos psicológicos. Neste enfoque, os itens não são construídos a partir de uma teoria; eles são coletados ou selecionados de um tal "pool of items" que parecem medir um dado construto e, em seguida, analisados estatisticamente para ver quais deles se salvam. Quer dizer, os itens são aqui simplesmente chutados; eles são selecionados simplesmente porque *parecem* medir o que quero medir.

Dentro da técnica de construção de instrumentos baseada na teoria dos traços latentes que estamos expondo, para se salvarem 20 itens no final de toda a elaboração e validação do instrumento, não é necessário iniciar com mais do que 10% de itens além dos 20 requeridos no instrumento final. Isto porque os itens incluídos no instrumento piloto são itens que possuem validade teórica real e não simplesmente *parecem* ter validade.

2.6 Análise Teórica dos Itens

Operacionalizado o construto através dos itens, estou diante da hipótese de que estes representam adequadamente o tal construto. Esta é a minha versão da hipótese a ser testada. Contudo, é importante avaliar esta minha hipótese contra a opinião de outros para me assegurar de que ela apresenta garantias de validade. Esta avaliação ou análise da hipótese (análise dos itens) é obviamente ainda teórica porque consiste simplesmente em pedir outras opiniões sobre minha hipótese, sendo que estes outros que a vão avaliar ainda não são uma amostra representativa da população para a qual construí o instrumento. Esta análise teórica é feita por juizes e ela comporta dois tipos distintos de juizes, segundo a análise incida sobre a compreensão dos itens (análise semântica) ou sobre a pertinência dos itens ao construto que representam (propriamente chamada de análise dos juizes). Assim, antes de partir para a validação final do instrumento piloto, este é submetido a uma análise teórica dos itens através da análise semântica e análise dos juizes.

2.6.1 Análise Semântica dos Itens

A análise semântica tem como objetivo precípuo verificar se todos os itens são compreensíveis para todos os membros da população à qual o instrumento se

destina. Nela duas preocupações são relevantes: 1) verificar se os itens são inteligíveis para o estrato mais baixo (de habilidade) da população meta e, por isso, a amostra para esta análise deve ser feita com este estrato; 2) para evitar deselegância na formulação dos itens, a análise semântica deverá ser feita também com uma amostra mais sofisticada (de maior habilidade) da população meta (para garantir a chamada 'validade aparente' do teste). Entende-se por estrato mais baixo aquele segmento da população meta que apresenta menor nível de habilidades. Assim, por exemplo, se meu teste se destina a uma população que congrega sujeitos do I grau de ensino até universitários, obviamente o estrato mais baixo neste contexto são os sujeitos do I grau e o mais sofisticado será representado pelos sujeitos de nível universitário. De qualquer forma, a dificuldade na compreensão dos itens não deve se constituir em fator complicador na resposta dos indivíduos, dado que não se quer medir a compreensão deles (a não ser, obviamente, que o teste queira medir precisamente isto), mas sim a magnitude do atributo a que os itens se referem. Que técnica utilizar para fazer esta análise? Há várias maneiras eficientes para tal tarefa, como por exemplo, aplicar o instrumento a uma amostra de uns 30 sujeitos da população meta e em seguida discutir com eles as dúvidas que os itens suscitarem. Entretanto, uma técnica que se tem mostrado das mais eficazes na avaliação da compreensão dos itens consiste em checá-los com pequenos grupos de sujeitos (3 ou 4) numa situação de "brainstorming". Esta técnica funciona da seguinte forma: Constituo um grupo de até 4 sujeitos, iniciando com sujeitos do estrato baixo da população meta, porque se supõe que se tal estrato compreende os itens, a fortiori o estrato mais sofisticado também os compreenderá. A este grupo apresento item por item, pedindo que ele seja reproduzido pelos membros do grupo. Se a reprodução do item não deixar nenhuma dúvida, o item é corretamente compreendido. Se surgirem divergências na reprodução do item ou se o pesquisador se aperceber que ele está sendo entendido diferentemente do que ele, pesquisador, julga que deveria ser entendido, este item tem problemas. Dada esta situação, o pesquisador então explica ao grupo o que ele pretendia dizer com tal item. Normalmente, neste caso, os próprios sujeitos do grupo irão sugerir como se deveria formular o item para expressar o que o pesquisador queria dizer com ele; e aí está o item reformulado como deve ser. Quantos grupos são necessários para proceder a esta análise semântica? Bem, itens que não ofereceram nenhuma dificuldade de compreensão em uma, no máximo duas, sessões, não necessitam de checagem ulterior. Itens que continuam apresentando dificuldades após, digamos, no máximo de cinco sessões, merecem ser simplesmente descartados. Em seguida a estas sessões, é importante pelo menos uma sessão de checagem dos itens com um grupo de sujeitos mais sofisticados. O objetivo desta verificação consiste em evitar que os itens se apresentem demasiadamente primitivos para estes sujeitos e assim perderem a validade aparente. É que os itens devem também dar a impressão de seriedade, como diz o ditado de que a mulher de César não somente deve ser honesta, mas deve também parecer honesta! (veja regra número 10 dos critérios de construção de itens).

2.6.2 Análise dos juizes

Esta análise é, às vezes, chamada de análise de conteúdo, mas propriamente deve ser chamada de análise de construto, dado que precisamente procura verificar a adequação da representação comportamental do(s) atributo(s) latente(s).

Na análise de conteúdo, os juizes devem ser peritos na área do construto, pois sua tarefa consiste em ajuizar se os itens estão se referindo ou não ao traço em questão. Uma tabela de dupla entrada, com os itens arrolados na margem esquerda e os traços no cabeçalho, serve para coletar esta informação. Uma concordância de, pelo menos, 80% entre os juizes pode servir de critério de decisão sobre a pertinência do item ao traço a que teoricamente se refere.

A técnica exige que se dê aos juizes duas tabelas: uma com as definições constitutivas dos construtos/fatores para os quais se criaram os itens e outra tabela de dupla entrada com os fatores e os itens, como na figura 3-3, onde são avaliados os itens que medem os dois fatores, compreensão verbal e fluência verbal, de raciocínio verbal. Normalmente, é necessária uma terceira tabela que elenca os itens, uma vez que a tabela de dupla entrada geralmente não comporta a expressão completa do conteúdo dos itens.

Fatores	Definição
Compreensão verbal	é a capacidade de...
Fluência verbal	é a capacidade de...

Itens	Compreensão verbal	Fluência Verbal
1	X	
2		X
3		X
...		
n		

Fig. 3-3. Tabelas para a análise dos itens pelos juizes

Com base nestas tabelas, a função dos juizes consiste em colocar um X para o item debaixo do fator ao qual o juiz julga o item se referir. Uma meia dúzia de juizes será suficiente para realizar esta tarefa. Itens que não atingirem uma concordância de aplicação aos fatores (cerca de 80%) obviamente apresentam problemas e seria o caso de descartá-los do instrumento piloto. Isto vale, contudo, se o construto para o qual estou construindo o teste apresentar fatores (particularmente quando forem em maior quantidade) que se supõem ou se sabe que não são correlacionados. Quando os fatores se supõem que sejam correlacionados, acontece que uma mesma tarefa (item) pode se referir, certamente com níveis de saturação diferente, mas de fato se referir simultaneamente a mais de um fator, o que implicaria que os juizes iriam mostrar alguma discordância quanto à aplicação do item a este ou a aquele fator. Neste caso, esta discordância deve ser considerada como concordância. Uma outra solução seria instruir os juizes a marcarem, para cada item, não o fator mas aqueles fatores aos quais o item se refere. Entretanto, com tal dica, você abre campo para muita divagação por parte dos juizes e você perde a utilidade prática desta análise. Seria melhor instruir os juizes para colocarem, *se possível*, cada item sob um fator somente.

Com o trabalho dos juizes ficam completados os procedimentos teóricos na construção do instrumento de medida, os quais comportaram a explicitação da teoria do(s) construto(s) envolvido(s), bem como a elaboração do instrumento piloto que constitui a representação comportamental destes mesmos construtos e que se põe como a hipótese a ser empiricamente testada (validação do instrumento), tarefa que será iniciada com os procedimentos que seguirão, os quais consistem em coletar informação empírica válida e submetê-la às análises estatísticas pertinentes em Psicometria, como veremos.

III. Procedimentos Experimentais

Os procedimentos envolvidos nesta etapa fazem apelo direto ao conteúdo da disciplina ensinada nas instituições universitárias sob o nome de Delineamento ou Planejamento de Pesquisa, cujo conhecimento é absolutamente necessário, vez que ela garante a tecnologia da coleta válida da informação empírica. Aqui serão, por isso, explicitados apenas alguns pontos desta tecnologia que têm mais a ver diretamente com o problema de elaboração de instrumentos psicológicos, mas o conhecimento aprofundado da citada disciplina é imprescindível.

Dois passos são salientados nestes procedimentos empíricos na validação do instrumento piloto: o planejamento da aplicação e a própria coleta da informação empírica, conforme detalha a figura 3-4.

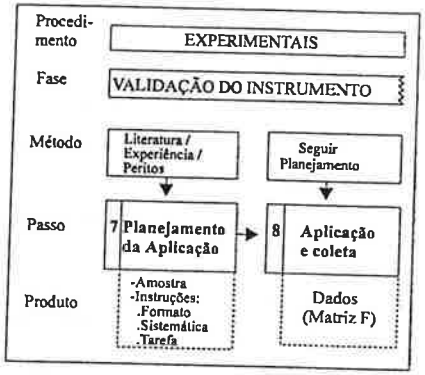


Fig.3-4. Procedimentos empíricos na elaboração de medida psicológica.

Com referência ao planejamento da aplicação do instrumento piloto, dois pontos são particularmente relevantes: a definição da amostra e das instruções de como aplicar o instrumento.

Quanto à *amostra*: um instrumento é tipicamente construído para um certo tipo de população. Esta, consequentemente, deve ser claramente definida e delimitada em termos de suas características específicas. Assim, é necessário se determinar para que faixa etária o instrumento foi construído, para que nível sócio-econômico, para que nível de escolaridade, etc. Enfim, é preciso dizer qual é o tipo de indivíduo, em termos de características bio-sócio-demográficas, que constitui a população meta do instrumento. E será desta população que sairá a amostra de sujeitos para a testagem da qualidade psicométrica do instrumento de medida. Obviamente, aqui se deve recorrer à teoria e técnicas de amostragem, ensinadas na disciplina Planejamento de Pesquisa ou similar. Salientamos aqui apenas alguns aspectos relevantes desta amostra para o caso específico de validação de instrumentos psicológicos. Como estamos elaborando um instrumento referente a construto, tipicamente a análise estatística a seguir utilizada para a análise dos dados será a análise fatorial e as análises multivariadas da TRI. Estas técnicas estatísticas fazem algumas exigências importantes dos dados, especificamente que eles produzam suficiente variância para que a análise seja consistente. Esta afirmação normalmente implica, pelo menos, que

as amostras utilizadas sejam grandes. Quanto grandes? Há duas dicas úteis para responder a esta pergunta. Primeiro, se eu estiver seguro de quantos fatores o meu instrumento mede (o que foi teoricamente definido quando se discutiu o passo da dimensionalidade do objeto psicológico que o instrumento iria medir), então a dica é de que a amostra deve conter um mínimo de 100 sujeitos por fator medido. Assim, se meu instrumento mede dois fatores, necessito de 200 sujeitos na minha amostra. Estamos supondo aqui que a população meta seja homogênea em relação ao traço latente que o instrumento mede. Se o traço varia dentro da população, não somente em termos de magnitude, o que é de se esperar, mas em termos de estrutura, isto é, ele se torna de fato um traço psicologicamente diferente para diferentes estratos da mesma população, então estamos falando não mais de um traço latente mas de dois ou mais. Neste caso, estamos assumindo que instrumentos diferentes são necessários para avaliar traços diferentes. Mas se o traço se mantém qualitativamente (em termos de sua estrutura conceitual, de sistema) o mesmo na população, então esta população é homogênea. Um exemplo: um teste de inteligência para adultos não inclui crianças na sua população, pois a inteligência da criança é qualitativamente diferente da dos adultos, segundo teorias (Piaget, Spearman, etc.) e dados empíricos. Assim, a amostra para validação de um teste de inteligência para adultos deve ser selecionado de uma população de adultos exclusivamente, que, neste sentido, se torna uma população homogênea.

Segunda dica: se se tiver dúvidas sérias quanto ao número de dimensões ou fatores que o instrumento mede, costuma-se dizer que são necessários para a amostra 10 sujeitos por cada item do instrumento. Assim, um instrumento com 100 itens demandaria 1.000 sujeitos. O que eqüivaleria a supor que o instrumento estivesse medindo cerca de 10 fatores. Este modo de pensar está mais ligado ao sistema positivista de construir instrumentos, no qual os itens não são construídos via teoria e sim "pescados" aleatoriamente e em seguida analisados via análise fatorial para ver quantos fatores está medindo. De qualquer forma, é uma dica ainda útil, quando dúvidas há com respeito ao número de fatores. Geralmente, entre 5 a 10 sujeitos por item do instrumento serão suficientes para responder à questão do tamanho da amostra, com a ressalva de que qualquer análise fatorial e da TRI com menos de 200 sujeitos dificilmente pode ser considerada adequada.

Quanto às *instruções*: Estas se referem aos contornos da tarefa do sujeito que vai responder ao instrumento. Aqui são definidas a sistemática de aplicação do instrumento, o formato em que ele se apresenta e o que o sujeito tem que fazer ao respondê-lo. No tocante à *sistemática*, serão definidas as condições de aplicação: será coletiva, individual; será preciso ou não aviso prévio aos testandos ou não; são necessários contatos prévios com diretores, chefes, etc. dos sujeitos, etc. Enfim, devo saber em que estou me metendo e quais são as dificuldades que vou encontrar ao querer aplicar o instrumento numa amostra definida de sujeitos, pois estes normalmente não estão gratuitamente disponíveis às minhas necessidades de pesquisador. Por isso, tenho que elaborar uma estratégia de convencimento, para os responsáveis dos sujeitos que entrarão na amostra, e uma estratégia operacional para poder viabilizar a aplicação do instrumento.

No referente ao formato do instrumento, deve-se decidir como a resposta do sujeito será dada para cada item. Aqui existe uma infinidade de formatos

possíveis, como, por exemplo, o da *escolha forçada* na qual dois itens são apresentados simultaneamente, sendo a tarefa do sujeito a de escolher um deles como mais apropriado, mais típico, ou mais o que seja, bastante comuns em testes de personalidade e mais ainda em testes de interesse; o das *múltiplas alternativas*, mais comuns em testes de aptidão, onde o sujeito deve escolher a alternativa correta; o das *escalas tipo Likert*, onde a cada item segue uma escala de pontos (de 2 a mais de 10) que exprimem a intensidade de acordo do sujeito com o que o item está afirmando. Este último formato é o mais utilizado no caso de testes de personalidade e escalas de atitudes. Todos estes e outros formatos apresentam vantagens e desvantagens. Por exemplo, o caso da escolha forçada, em testes de atitudes e personalidade, parece ser a maneira mais fácil de responder, pois o sujeito tem melhores condições de escolher entre duas alternativas do que dar uma resposta absoluta como é o caso nas escalas de Likert. Contudo, dois problemas graves existem com este formato de escolha forçada: primeiro, se você vai comparar os itens do instrumento dois a dois, o instrumento se torna muito rapidamente de um comprimento incontrolável. Por exemplo, um teste com apenas 10 itens, terá $\frac{n(n-1)}{n}$ questões, isto é, $\frac{10(10-1)}{10} = 45$ questões e um de 100 itens terá 4.950! Além desta dificuldade, existe o problema da chamada desejabilidade social, a saber, os dois itens que estão sendo comparados devem possuir mais o menos o mesmo nível de atratividade, do contrário a própria questão já está dando a resposta ao sujeito se um dos itens da questão é socialmente desejável e o outro indesejável, como por exemplo escolher entre "A - sou uma pessoa simpática" e "B - sou uma pessoa fraca". Neste caso, a maioria das pessoas iria escolher a alternativa A. Certo?

No caso do formato de mútipla escolha, existem os problemas do número de alternativas e da qualidade das alternativas. Primeiramente, como se trata de respostas certas e erradas, apenas uma das alternativas será a correta. Mas, quando o sujeito não sabe a resposta correta, ele tem a chance de "chutar" e acertar por acaso; e isto é um problema, que é tanto mais grave quanto menor for o número de alternativas. Por exemplo, num item com 2 alternativas, o sujeito tem a chance de acertar por acaso em 50% das vezes, ao passo que num item com 5 alternativas esta chance cai para 20%, mas ainda não é zero. Então, deve-se ter maior do que menor número de alternativas para diminuir o acerto aleatório. Mas, fazendo isto, você torna o teste cada vez mais difícil de construir, porque não é tarefa fácil inventar alternativas, uma vez que estas devem de fato se apresentar como alternativas plausíveis e atrativas (e este é o segundo problema), isto é, elas devem ter alguma aparência de serem respostas corretas, do contrário não são alternativas. Assim se você constrói o seguinte item:

A camada mais externa da pele se chama:

a) epiderme
b) paquiderme
c) dermatologia
d) epidemia

é claro que b, c, d não constituem alternativas plausíveis ou sérias.

Quanto às escalas tipo Likert, pergunta-se freqüentemente qual é o número ideal de pontos que a escala de resposta deve ter e qual o formato ideal da escala.

Com respeito ao formato das escalas: Existem os mais variados modos de apresentar estas escalas, mas que finalmente se reduzem a escalas verbais, numéricas ou escalas gráficas, sendo estas últimas normalmente ancoradas, ou combinação das três.

Vejamos:

Escala verbal:	Sim			Em dúvida			Não
Escala numérica:	1	2	3	4	5	6	7
Escala gráfica:	----- :	----- :	----- :	----- :	----- :	----- :	-----
Escala gráfica ancorada	acordo						desacordo
	----------						----------
Escala numérica e gráfica:	1 ----- :	2 ----- :	3 ----- :	4 ----- :	5 ----- :	6 ----- :	7 -----

Estes e outros tipos de formatos não parecem ter maior impacto sobre a resposta do sujeito; de sorte que o formato da escala depende mais do gosto pessoal do pesquisador do que qualquer outra razão técnica. Pessoalmente acho que quanto mais leve a escala, melhor; assim, a escala numérica e gráfica me parece muito pesada; mas sua opinião é tão boa quanto a minha.

Quanto ao número de pontos: normalmente as afirmações ou itens são respondidos numa escala de 3 ou mais pontos, isto é, o sujeito tem que dizer se concorda, está em dúvida ou discorda com o que a frase afirma sobre o objeto psicológico. O número de pontos na escala de resposta varia de 3 a mais de 10, sendo as mais utilizadas as escalas de 5 e 7 pontos. O número de pontos utilizados nas escalas Likert parece, novamente, ser algo irrelevante. Na pesquisa de Matell e Jacoby (1972), foram utilizadas escalas com 2 até 19 pontos. Com exceção das escalas de 2 e 3 pontos (por oferecerem poucos graus de liberdade), em todas as outras a porcentagem de uso dos pontos e o tempo de resposta não foram afetados de modo significativo. Outros estudos já haviam descoberto que o número de pontos da escala, bem como a existência ou não de um ponto neutro, não afeta a consistência interna da escala Likert (Bendig, 1954; Komorita, 1963; Matell & Jacoby, 1971), nem a estabilidade teste-reteste (Jones, 1968; Van der Veer, Howard & Austria, 1970; Goldsamt, 1971; Matell & Jacoby, 1971) e nem a validade concorrente e preditiva (Matell & Jacoby, 1971, 1972).

As *instruções* que acompanham o instrumento têm a função única de tornar a tarefa do respondente inambígua. Consequentemente, elas devem poder deixar absolutamente claro o que o sujeito tem que fazer para responder corretamente o teste e, por isso, elas devem ser avaliadas na análise semântica. Algumas precauções: as instruções devem informar em termos gerais sobre que é o teste; elas devem ser o mais curtas possível, sem sacrificar a compreensão da tarefa por parte de todos os sujeitos da população meta; elas devem, tipicamente, conter um ou mais exemplos de como os itens devem ser respondidos; elas devem pôr o sujeito num estado psicológico livre de tensão e ansiedade.

Finalmente, no que se refere à própria coleta da informação (passo 8), deve-se seguir todas as precauções que se exigem em qualquer aplicação de instrumentos psicológicos, a saber, pôr os sujeitos num ambiente condizente e livre de distrações e de tensão, o aplicador ser competente para a tarefa etc.

IV. Procedimentos Analíticos

Esta parte da elaboração de instrumentos psicológicos (veja Fig. 3-5) é aquela que mais atemoriza os psicólogos, dada a sua sofisticação estatística. Ela comporta igualmente a parte mais volumosa de qualquer livro sobre Psicometria. Entretanto, o conhecimento da Estatística e da Psicometria não são aqui substituíveis. Felizmente, o psicólogo pode apelar neste particular para a ajuda de estatísticos ou de psicometristas. A sofisticação nesta área é tão grande que não é possível ser exposta neste capítulo. Para tanto, são recomendadas as obras que em seguida serão citadas, sendo a exposição de conteúdo neste capítulo apenas exemplificativa.

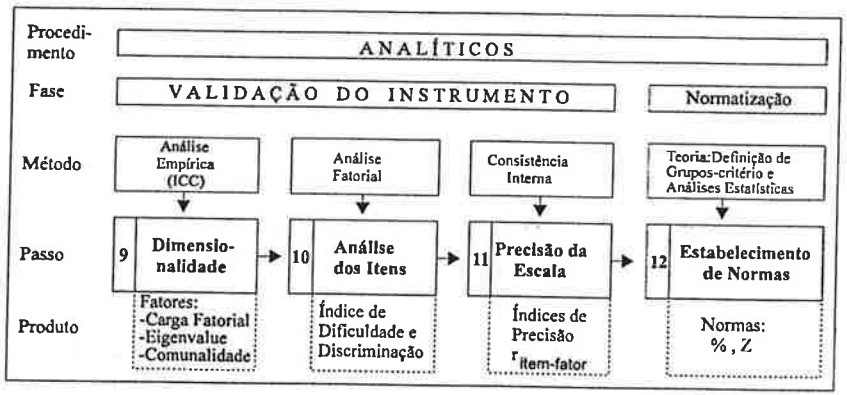


Fig.3-5. Procedimentos analíticos na elaboração da medida psicológica.

4.1 Algumas Obras Básicas de Análise Psicométrica

Anastasi, A. (1988). *Psychological testing.* 6th ed. New York: Macmillan Publ. Co.

Cronbach, L.J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.

Gulliksen, H. (1987). *Theory of mental tests.* Hillsdale, NJ: Erlbaum.

Hambleton, R.K., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of Item Response Theory.* Beverly Hills, CA: SAGE.

Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and applications.* Norwell, MA: Kluwer Nijhoff.

Hambleton, R.K. & Zaal, J.N. (eds. - 1991). *Advances in educational and Psychological testing: Theory and applications.* Boston, MA: Kluwer Academic Publishers.

Harman, H.H. (1967). *Modern factor analysis*. Chicago, IL: University of Chicago Press.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). *Item Response Theory. Applications to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

Lord, F.M. (1980). *Applications of Item Response Theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Muñiz, J. (1990). *Teoría de respuesta a los ítems*. Madrid: Pirámide.

Muñiz, J. (1992). *Teoría clássica de los tests*. Madrid: Ediciones Pirámide, S.A.

Muñiz, J. (1996). *Psicometría*. Madrid: Univérsitas.

Nunnally, J.C., Jr. (1978). *Psychometric theory*. New York: McGraw-Hill.

Pasquali, L. (Org. - 1996). *Teoria e métodos de medida em ciências do comportamento*. Brasília: INEP.

Pasquali, L. (1998). *Psicometria: Teoria e Aplicações*. Brasília: Editora UnB.

Santisteban, C. (1990). *Psicometría*. Madrid: Norma.

Thorndike, R.L. (1982). *Applied psychometrics*. Boston: Hougton Mifflin.

Yela, M. (1987). *Introducción a la teoría de los tests*. Madrid: Facultad de Psicología, Universidad Complutense.

4.2 Algumas Anotações Sobre os Procedimentos Analíticos

1) A Dimensionalidade do Instrumento (Validade)

As análises estatísticas que se fazem de um instrumento psicológico, no seu todo e em cada item individual, fazem a suposição de que o instrumento seja unidimensional. Isto implica que todos os itens do instrumento estejam medindo um e o mesmo construto. De sorte que, estando o instrumento medindo mais de um fator, as análises estatísticas devem ser feitas independentemente para cada fator. Como inicialmente ainda não se sabe se o instrumento que acaba de ser construído e aplicado é ou não unidimensional, a primeira análise que se impõe sobre os dados empíricos coletados é a verificação desta unidimensionalidade. Tipicamente, necessita-se proceder a uma análise fatorial para definir a *dimensionalidade* do instrumento. Esta análise vai determinar quantos fatores meu instrumento está de fato medindo. Esta exigência pode parecer um tanto frustrante, uma vez que, se construí o instrumento para medir um fator somente, por exemplo, então não posso supor que esteja medindo somente este fator para o qual construí o instrumento em primeiro lugar? É bom se lembrar aqui que o instrumento constitui uma hipótese; mas, agora estamos verificando esta hipótese empiricamente, então é necessário se demonstrar e não somente supor que o instrumento de fato mede um único fator ou quantos e quais fatores ele está medindo. Aliás, esta análise fatorial constitui a demonstração da própria validade do instrumento e representa igualmente a análise preliminar dos próprios itens.

A *análise fatorial* (veja Pasquali, 1999) produz resultados importantes com os quais se pode tomar decisões sobre a qualidade dos itens, bem como do

instrumento no seu todo. Na verdade, ela mostra o que o instrumento está medindo, isto é, os fatores, bem como os itens que compõem cada fator. Ela produz, para cada item, a carga fatorial (saturação) deste no fator e esta carga fatorial indica a covariância entre o fator e o item. Isto quer dizer, a carga fatorial mostra quanto porcento existe de parentesco (covariância) entre o item e o fator, de sorte que quanto mais próximo de 100% de covariância item-fator, melhor será o item, pois ele assim se constitui num excelente representante comportamental do fator (do traço latente). Qual é o montante de covariância entre o item e o fator necessário para se dizer que o item é um bom representante deste? As cargas fatoriais são expressas similarmente aos índices de correlação e, portanto, podem ir de -1,00 a +1,00. Uma carga de 0,00 significa que não há nenhuma relação entre o item e o fator; neste caso, o item seria uma representação comportamental totalmente equivocada do fator. Então, que nível de magnitude de carga deve o item apresentar para ser um bom representante do fator? Costuma-se apontar o valor 0,30 (positivo ou negativo) como sendo uma carga mínima necessária para o item ser um representante útil do fator. Obviamente, quanto maior de 0,30 for a carga, melhor o item. Uma carga de 0,30 indica que há uma covariância de cerca de 10% ($0{,}30^2 = 0{,}09$) entre o item e o fator, o que já pode ser considerado não negligível, embora não seja lá grande coisa. Obviamente se todos os itens de um fator apresentam cargas fatoriais em torno de 0,30, este fator está muito mal representado, porque se esperam cargas bem maiores (acima de 0,50) para se dizer que o fator foi bem representado comportamentalmente. Você vê, então, que as cargas fatoriais falam tanto da qualidade de cada item, bem como do conjunto deles, isto é, do próprio fator. Assim, se você construiu 25 itens para representar o traço latente e, destes 25 itens, 20 apresentam cargas acima de 0,50 e 5 apresentam cargas em torno de 0,30, você irá já eliminar estes últimos 5 itens e trabalhar somente com os 20 itens que apresentaram cargas fatoriais respeitáveis. Veja o exemplo (fictício) da tabela 3-1.

A Tabela 3-1 exemplifica uma típica matriz fatorial com as informações essenciais sobre os itens e os fatores. Nela se vê que dos 20 itens, 9 (com cargas fatoriais em negrito) representam o fator 1, pois possuem cargas fatoriais altas neste fator e praticamente cargas nulas no fator 2; ao contrário, os 10 últimos itens possuem cargas fortes no fator 2 e quase nada no fator 1. O item 10 não possui carga expressiva em nenhum dos dois fatores e será, por isso, descartado do teste. Observe que as cargas fatoriais podem ser tanto positivas quanto negativas e, assim mesmo, pertencerem ao mesmo fator, conquanto que elas sejam altas. É que o fato delas serem positivas e negativas no mesmo fator apenas indicam que um item está expresando o pólo positivo e o outro o pólo negativo do fator, como, por exemplo, estes dois itens "gosto de meus pais" e "detesto meus pais", ambos se referem à questão da afiliação, apenas que o primeiro item expressa o pólo positivo da afiliação e o segundo, o pólo negativo.

Assim, o teste mede dois fatores, um com 9 itens e o outro com 10, mostrando-se um item (o número 10) uma representação equivocada tanto do fator 1 quanto do fator 2. Os dois fatores explicam 47,73 % (= 23,07 + 24,66) da variância total do teste, sendo o restante da variância irrelevante ao conteúdo que o teste mede (como, erros de medida e peculiaridades específicas dos itens). O h^2 representa a comuniladade que cada item possui com os dois fatores, e mostra a covariância de

Tabela 3-1. Cargas fatoriais de 20 itens em dois fatores

Item	Fator 1	Fator 2	h^2
1	.80	.10	.65
2	.78	-.05	.61
3	.78	.20	.65
4	.70	.15	.51
5	.65	.08	.43
6	.64	.12	.42
7	-.64	-.10	.42
8	.60	.03	.36
9	-.60	-.23	.41
10	-.25	.19	.10
11	.30	-.83	.78
12	.21	-.80	.68
13	.04	-.78	.61
14	.16	-.70	.52
15	-.12	.70	.50
16	.09	.66	.44
17	-.00	-.65	.42
18	.12	-.63	.41
19	-.03	.56	.31
20	.21	-.50	.29
Eigenvalue	4.614	4.932	
% Var. total	23.07	24.66	
% Var. comum	48.33	51.67	

item com os fatores e, por conseguinte, o tanto que o item tem a ver com os fatores. Assim, para o item 1, o h^2 é 0,65, isto é, este item possui 65 % de covariância (parentesco) com os dois fatores, sendo 0,64 % ($0,80^2$) com o fator 1 e apenas 1 % ($0,10^2$) com o fator 2; donde se deduz que o item 1 é uma excelente representação comportamental do fator 1 e nada do fator 2.

Nesta questão da validade do instrumento, outras técnicas são utilizadas além da análise fatorial, tais como, a técnica da *validação convergente-discriminante* (Campbell & Fiske, 1967); a utilização da *idade* como critério para a validação de construto de um teste quando este mede traços que são intrinsecamente dependentes de mudanças no desenvolvimento cognitivo/afetivo dos indivíduos, como é o caso, por exemplo, na teoria piagetiana do desenvolvimento dos processos cognitivos e da teoria de Spearman sobre a inteligência; a *correlação com outros testes* que meçam o mesmo traço do meu novo instrumento e o uso da *intervenção experimental* (veja Pasquali, 1998).

2) A Análise Empírica dos Itens

Os itens que se mostraram ser representantes satisfatórios do traço latente que o instrumento mede (no caso da tabela 3-1 seriam os 9 itens para o fator1 e os 10 do fator2) devem ser submetidos a análises individuais ulteriores, com o

objetivo de verificar outras características que eles devem apresentar dentro de um mesmo instrumento, além de serem legítimos representantes do traço latente. Estas características dos itens devem ser analisadas dentro de cada fator (os 9 itens no fator1 do nosso exemplo, e os 10 no fator2) e normalmente se reduzem a duas: a dificuldade e a discriminação. A dificuldade do item diz respeito à magnitude do traço latente que o sujeito deve possuir para poder acertar (testes de aptidão) ou aceitar (testes de personalidade) o item. Assim, quanto maior for a magnitude do traço latente exigida para acertar ou aceitar o item, mais difícil este é dito ser. A discriminação do item diz respeito ao fato dele poder diferenciar sujeitos que possuem magnitudes diferentes do mesmo traço latente. Assim, quanto mais próximas forem as magnitudes do traço que o item puder diferenciar, mais discriminativo ele será.

A Psicometria tradicional fazia análises estatísticas para determinar estes dois parâmetros psicométricos dos itens de uma forma que podem ser hoje consideradas obsoletas diante dos avanços da Psicometria moderna da Teoria de Resposta ao Item (TRI). A TRI introduziu técnicas nesta área da análise dos itens que, embora complicadas, devem ser as utilizadas neste passo da elaboração de qualquer instrumento psicológico (veja Hambleton, Swaminathan, & Rogers, 1991; Muñiz, 1990). Um exemplo ajudará a entender estes procedimentos (veja figura 3-6).

Primeiramente, deve-se atentar a que existem vários modelos matemáticos envolvidos na TRI. Na verdade, há três deles principais, dependendo do número de parâmetros que pretendem avaliar dos itens. Os parâmetros em questão são a dificuldade, a discriminação e a resposta aleatória (ou melhor, a resposta correta dada ao acaso). Assim, temos os modelos logísticos de 1, 2 ou 3 parâmetros.

Todos os modelos trabalham com traços latentes, isto é, teorizam sobre as estruturas latentes. Entendem os sistemas psicológicos latentes como possuindo dimensões, isto é, propriedades de diferentes magnitudes ou mensuráveis. Por isso, esta teoria também é conhecida como a teoria do traço latente ou a teoria da curva característica do item ("item characteristic curve" - ICC), pelo fato de produzir para cada item uma ogiva característica dele. A teoria supõe que o sujeito possui um certo nível de magnitude do traço latente, designado por theta (θ), o qual é determinado através da análise das respostas dos sujeitos por meio de diversas funções matemáticas. A função do modelo completo de três parâmetros é:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

A probabilidade de resposta correta, que define a posição (θ) do indivíduo no traço medido, é função de três parâmetros: "a" corresponde ao índice de discriminação do item e é determinado pela inclinação da curva no ponto de inflexão; "b" é o parâmetro da dificuldade/preferência e é expresso pelo valor no eixo dos X no ponto de inflexão da curva; "c" é o parâmetro que determina as respostas acertadas/preferidas por acaso, sendo o D uma constante usualmente com valor 1.7.

Os três modelos de TRI mais conhecidos são os seguintes: 1) o modelo logístico de um parâmetro ou o modelo Rasch (1960). Rasch faz a suposição de que os itens possuem o mesmo nível de discriminação e que não há respostas dadas ao acaso, ficando como parâmetro a ser avaliado somente a dificuldade dos itens. 2)

O modelo logístico de dois parâmetros (Birnbaum, 1968), que avalia a dificuldade e a discriminação dos itens, assumindo que não hajam respostas dadas ao acaso. 3) O modelo de três parâmetros de Lord (1980) no qual os três parâmetros dos itens são avaliados.

Exemplificando com o modelo de Lord: Os valores θ são expressos em coordenadas cartesianas, tendo na ordenada a probabilidade de resposta correta, isto é, o $P_i(\theta)$, e na abcissa o traço latente, o próprio θ. Este procedimento produz, para cada item, uma ogiva, chamada de curva característica do item ("item characteristic curve" ou ICC), como na Figura 3-6.

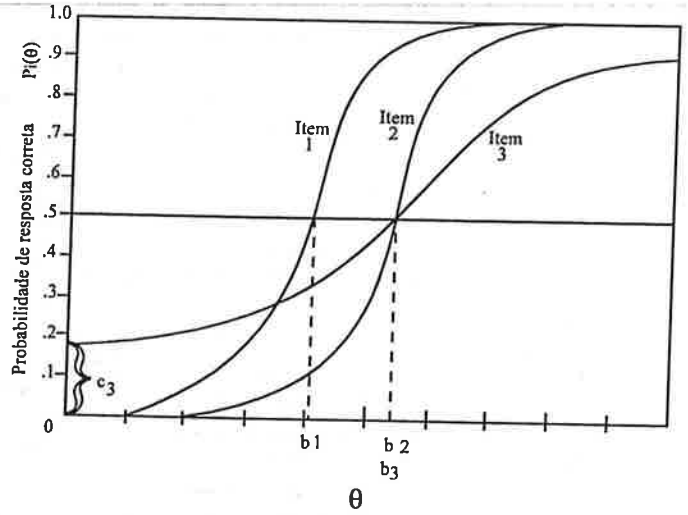


Fig.3-6. ICC para três itens.

Na ilustração da figura 3-6, os três parâmetros aparecem nas seguintes posições: o "a" é representado pela inclinação da curva na altura do ponto de inflexão, isto é, onde a curva corta a linha que representa a probabilidade .50 de resposta correta (50%); quanto mais íngreme esta curvatura, isto é, quanto mais próxima de um ângulo de incidência de 90°, mais discriminativo é o item. O "b" é representado pela distância na linha dos X (abcissa), que corresponde ao ponto determinado pela perpendicular que vem do ponto de inflexão da curva. O "c" é definido pela assíntota inferior da curva; quando esta assíntota não atinge a abcissa, há respostas dadas ao acaso e o tamanho destas respostas é definido pela distância que vai do ponto 0 na abcissa até o ponto onde a ogiva corta a ordenada; por exemplo, o item 3 tem cerca de 18% de resposta ao acaso. Vê-se também nesta figura 3-6 que os itens 1 e 2 são mais discriminativos do que o item 3; igualmente que os itens 2 e 3 possuem o mesmo nível de dificuldade (mais ou menos 0,4 sigmas acima da média 0) e que o item 1 é o mais fácil dos três, com um índice de mais ou menos -0,70 sigmas da média. Os dados oferecidos pela TRI são algebricamente expressos numa tabela como a que segue (onde aparecem os dados dos três itens do exemplo da figura 3-6):

Item	Parâmetros		
	a	b	c
1	1,50	-0,70	0,00
2	1,50	0,40	0,00
3	0,40	0,40	0,18
...	...	...	...

Nível ideal de dificuldade dos itens. Pode-se perguntar, ainda, se existe um nível ideal de dificuldade para os itens de uma escala ou teste. Esta pergunta está relacionada com os critérios 11 e 12 (amplitude e equilíbrio dos itens no instrumento) das regras de construção dos itens. A resposta a esta indagação depende da finalidade do teste. Se se deseja um teste para selecionar os melhores ou para determinar se um patamar 'x' de conhecimento foi atingido (como nos testes educacionais de referência a critério), então os itens devem todos apresentar o nível de dificuldade do patamar que se quer como critério de seleção ou acima dele. Assim, se se deseja selecionar somente os 30% melhores candidatos, os índices de dificuldade dos itens devem ser em torno de 30% (p = 0,30) ou menos, isto é, somente 30% dos sujeitos devem ter a probabilidade de acertar os itens. De fato, neste caso, existe o interesse em apenas discriminar entre sujeitos de alta aptidão, sendo sem interesse itens que apenas discriminassem sujeitos de menor aptidão.

Se, entretanto, o interesse consiste em avaliar a magnitude diferencial dos traços nos sujeitos de uma população, como geralmente é o caso em testes referentes a construto, então uma distribuição mais equilibrada dos itens em termos de dificuldade é requerida. Neste caso, o interesse se centra sobre o poder de um teste discriminar diferentes níveis de habilidades nos sujeitos e, por conseguinte, os itens devem poder avaliar tanto os que possuem pouca quanto muita habilidade. Entretanto, é bom saber que itens que todos os sujeitos acertam ou aceitam e itens que ninguém acerta ou não aceitam são itens inúteis para fins de diferenciar indivíduos; de fato, tais itens não trazem nenhuma informação. Os itens que trazem maior informação são aqueles cujo índice de dificuldade se situa em torno de 50%, isto é, no valor 0 da escala dos sigmas, pois neste caso 50% dos sujeitos acertam e 50% erram, resultando 50x50 = 2.500 comparações possíveis, ao passo que um item com dificuldade 30% teria 70% de erros e 30% de acertos, resultando num nível de 30x70 = 2.100 bits de informação. Obviamente, um item com dificuldade 100% ou 0% produzirá zero informação. Deve-se concluir daí que todos os itens de um teste devam ter dificuldade 50%? Embora grande parte dos itens deva apresentar tal índice de dificuldade, nem todos o deverão, pois que assim poder-se-ia discriminar apenas dois níveis da magnitude do traço medido, dado que itens com o mesmo nível de dificuldade terão altas intercorrelações, determinadas pela circunstância de que serão os mesmos sujeitos que sempre acertam ou sempre erram os itens todos. Isto vale dizer que a dificuldade média dos itens do teste deve ser em torno de p = 0,50. Haveria, então, uma distribuição mais adequada dos itens de um teste em termos de dificuldade? Considerando que eles devem cobrir toda a extensão de magnitude do traço e que os itens de dificuldade 50% são os que produzem maior informação, pode-se sugerir que uma distribuição dos mesmos mais ou menos dentro de uma curva normal seria o ideal. Assim, se considerarmos a amplitude de um atributo ou traço numa escala de

100 pontos, podemos dividi-la em cinco níveis de magnitudes: 0 a 20 (sigma ≤ -1,28), 20 a 40 (sigma entre -1,28 e -0,52), 40 a 60 (sigma entre -0,52 e +0,52), 60 a 80 (sigma entre 0,52 e 1,28) e 80 a 100 (sigma ≥ 1,28), distribuindo os itens assim: 10% deles em cada uma das duas faixas extremas, 20% em cada uma das duas faixas seguintes e 40% na faixa média (vide Figura 3-7).

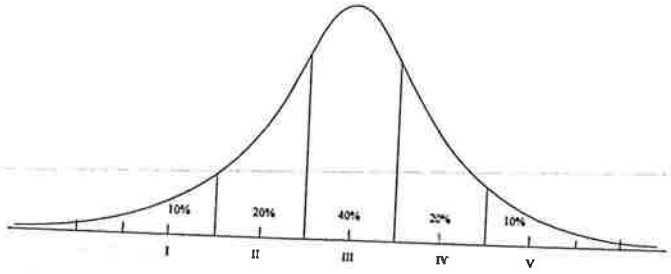


Fig. 3-7. Distribuição percentual dos itens em 5 faixas de dificuldade.

Essa discussão sobre a dificuldade ideal dos itens faz mais sentido dentro da teoria clássica dos testes. A TRI tem maneiras bem mais condizentes e apropriadas para fazer esta análise através do uso do índice de informação do item e do teste. Trabalhar com este índice é bem mais complexo, mas existem softwares apropriados em abundância no mercado para auxiliar nesta tarefa. Ademais, para podér fazer uso inteligente de tal procedimento é necessário um conhecimento razoável da TRI. Por isso, o leitor deve se aprofundar no estudo de algum dos livros acima citados sobre a TRI.

3) Fidedignidade do Instrumento

O problema que se enquadra sob o conceito de fidedignidade vem relatado sob uma série de outras expressões, como: precisão, fidedignidade, constância, consistência interna, confiabilidade, estabilidade, confiança, homogeneidade. As mais genéricas e, por isso, as mais utilizadas são as expressões precisão e fidedignidade.

Estas diferentes expressões mostram a variabilidade de conceitos que precisão assume, dependendo do aspecto que este parâmetro quer salientar do teste. Na verdade, fidedignidade cobre aspectos diferentes de um teste, mas todos eles se referem a quanto os escores de um sujeito se mantêm idênticos em ocasiões diferentes; por exemplo, os escores obtidos num tempo 1 e num tempo 2 para os mesmos sujeitos. Esta ocorrência (identidade dos escores) evidentemente supõe que o traço que o teste mede se mantenha constante sobre estas diferentes ocasiões, como é suposto ser o caso, por exemplo, na maioria dos traços de personalidade e de aptidão. Não seria o caso num teste de humor, porque este traço por natureza varia de momento para outro, e um teste válido de humor produziria escores necessariamente diferentes. Assim, o conceito de fidedignidade, na verdade, se refere ao quanto o escore obtido no teste se aproxima do escore verdadeiro do sujeito num traço qualquer; isto é, a fidedignidade de um teste está intimamente ligada ao conceito da variância erro, sendo este definido como a variabilidade nos escores produzida por

fatores estranhos ao construto. Aparece, assim, claro que a fidedignidade de um teste depende da questão do erro da medida, especificamente do erro produzido pelo próprio instrumento: quanto o escore produzido pelo teste se distancia do escore verdadeiro do sujeito no traço em questão, isto é, a valor theta individual na TRI.

Para melhor conceber esta problemática, é preciso se referir à variância verdadeira e variância erro. Um procedimento de medida qualquer, por exemplo os escores em um teste, produz uma variabilidade nos resultados, que, em parte é provocada pelas diferenças no próprio traço medido entre diferentes sujeitos, parte pela imprecisão do próprio instrumento e parte, ainda, por uma série de outros fatores aleatórios. A fidedignidade da medida depende do tamanho da variância erro, que é precisamente a variabilidade nos resultados provocada por estes fatores aleatórios e pela imprecisão do instrumento. Expressa mais positivamente, a fidedignidade de um instrumento diz respeito ao montante de variância verdadeira que ele produz vis-à-vis a variância erro, isto é, quanto maior a variância verdadeira e menor a variância erro, mais fidedigno o instrumento: um escore preciso é um escore que se aproxima do valor verdadeiro, expresso estatisticamente pelo erro padrão da medida (tratado mais adiante).

A definição estatística da fidedignidade é feita através da correlação entre escores de duas situações produzidos pelo mesmo teste. Se o teste é preciso, esta correlação deve ser, não somente significativa, mas se aproximar da unidade (cerca de 0,90). De fato, uma correlação de 0,70, por exemplo, expressaria uma comunalidade de apenas 49% entre as duas situações provocadas pelo mesmo teste nos mesmos sujeitos. Neste caso, a variância comum, digamos a variância verdadeira, seria menor que a variância erro, demonstrando que o teste não produz resultados fidedignos, isto é, o teste não possui precisão. Esta correlação, no caso do parâmetro de fidedignidade ou precisão, é referida como o coeficiente de precisão ou de fidedignidade.

Dependendo da técnica utilizada para cômputo da precisão de um teste, surgem vários tipos de precisão: teste-reteste, formas paralelas, consistência interna.

(1) A precisão *teste-reteste* consiste em calcular a correlação entre as distribuições de escores obtidos num mesmo teste pelos mesmos sujeitos em duas ocasiões diferentes de tempo. A correlação de 1,00 seria obtida se não houvesse variância erro provocada pelo teste ou outros fatores aleatórios, como fatores não controlados nos sujeitos ou na situação de testagem. Quanto mais longo o período de tempo entre a primeira e a segunda testagem, mais chances haverá de fatores aleatórios ocorrerem, diminuindo o coeficiente de precisão. Este intervalo de tempo permite a ação dos fatores mencionados por Campbell e Stanley (1963) sob o tema de fontes de erro devido à história, maturação, retestagem e às interações entre estes fatores, bem como ao próprio instrumento. Por isso, vêem-se as graves dificuldades que apresenta este tipo de análise da fidedignidade de um teste; particularmente grave aparece aqui a questão da maturação, isto é, se o próprio traço matura (se desenvolve, modifica), esta análise da precisão torna-se errônea, dada sobretudo a eventualidade de que a maturação do traço se processe diferencialmente para os diversos sujeitos testados. Além disso, e particularmente em testes de aptidão, a testagem constitui um treinamento, e provavel-

mente diferencial, para os sujeitos, o que provocará diferenças na retestagem entre os mesmos, reduzindo novamente o coeficiente de precisão do teste. Para contornar estas dificuldades, outros tipos de análises foram elaboradas, como a das formas alternativas ou análise da consistência interna.

(2) Na precisão de *formas alternativas*, os sujeitos respondem a duas formas paralelas do mesmo teste e a correlação entre as duas distribuições de escores constitui o coeficiente de precisão do teste. A condição necessária para que esta análise seja válida se situa na demonstração de que as amostras de conteúdo (de itens) em ambas as formas sejam equivalentes, isto é, que os itens possuam níveis equivalentes de dificuldade e de discriminação em ambas. Estes parâmetros podem ser facilmente verificados através da TRI. Há, contudo, algumas dificuldades neste tipo de análise: as duas formas são aplicadas em sucessão imediata, não eliminando assim totalmente o efeito do intervalo de tempo, resultando na possível introdução de efeitos da história e do treinamento (prática) obtido ao responder à primeira das formas alternativas; aparece facilmente um efeito repetitório, dado que os itens de ambas as formas são similares, produzindo efeitos motivacionais negativos no respondente. Além disso, não é tarefa fácil construir formas alternativas, quando a construção de um só teste já é uma tarefa dispendiosa, razão pela qual poucos testes aparecem no mercado com formas alternativas.

(3) A precisão da *consistência interna* é viabilizada através de várias técnicas estatísticas que visam verificar a homogeneidade da amostra de itens do teste, ou seja, a consistência interna do teste. As técnicas mais utilizadas são: duas metades, Kuder-Richardson e alfa de Cronbach. Todas elas exigem aplicação do teste em apenas uma única ocasião, evitando totalmente a questão da constância temporal.

No caso da precisão das *duas metades*, os sujeitos respondem a um único teste numa única ocasião. O teste é dividido em duas partes equivalentes e a correlação é calculada entre os escores obtidos nas duas metades. Não é importante como o teste é dividido em duas metades, conquanto que estas sejam equivalentes. Na prática, contudo, as duas formas mais normalmente utilizadas são a divisão do teste em primeira metade e segunda metade ou em itens pares e itens ímpares. Para efetuar esta análise, de fato o teste não precisa ser homogêneo, isto é, no qual todos os itens medem o mesmo traço (por exemplo, itens somente verbais ou numéricos); o que é fundamental é que as duas metades emparelhem itens homogêneos: verbal com verbal, numérico com numérico, etc.

Neste tipo de precisão, é preciso notar que o cálculo da correlação se baseia somente na metade do teste. Assim, num teste de 100 itens, a correlação se basearia somente em 50 itens. Como o número de itens afeta o tamanho do coeficiente de correlação, é preciso corrigir este coeficiente para que leve em consideração a extensão total do teste e, assim, produzir um coeficiente de precisão mais justo para o teste. Esta correção é feita através da fórmula de Spearman-Brown:

$$r_{tt} = \frac{nr_{12}}{1+r_{12}}$$

onde, r_{tt} é o coeficiente de precisão calculado, r_{12} é o coeficiente de correlação entre as duas metades do teste e *n* é o número de vezes em que o teste foi dividido. Assim, um teste dividido em duas metades, o *n* será 2, porque ele deve ser aumentado 2 vezes para se obter a forma total do teste.

A técnica de *Kuder-Richardson* (Kuder & Richardson, 1937) para verificar a fidedignidade de um teste se baseia na análise de cada item individual do teste. Os autores desenvolveram várias fórmulas, sendo a mais utilizada a fórmula 20, que segue:

$$r_{tt} = \left(\frac{n}{n-1}\right)\frac{DP_t^2 - \sum pq}{DP_t^2}$$

onde, r_{tt} é o coeficiente de precisão do teste,
n o número de itens do teste,
DP_t^2 o desvio padrão dos escores totais do teste e
$\sum pq$ é o somatório do produto da proporção de sujeitos que passaram (p) e dos que não passaram (q) cada item.

Cronbach (1951) mostrou que esta técnica produz um coeficiente de precisão do teste que corresponde à média dos coeficientes de todas as metades em que o teste possa ser dividido, mas somente quando se utiliza a fórmula de Rulon (1939), que trabalha com as variâncias das diferenças entre as duas metades, e não a simples correlação com a correção de Spearman-Brown, segundo observaram Novick e Lewis (1967). Esta equivalência de coeficientes, contudo, ocorre em testes homogêneos, porque nos testes heterogêneos os coeficientes de Kuder-Richardson são normalmente menores, dado que esta técnica não trabalha com diferenças entre pares de itens e sim com a variância de todos os itens.

O próprio Cronbach (1951) desenvolveu um técnica geral para estabelecer a fidedignidade dos testes, o *Alfa de Cronbach*. Esta constitui uma extensão da de Kuder-Richardson. Esta última é aplicável somente quando a resposta ao item é dicotômica: certo e errado, por exemplo. Entretanto, quando a resposta ao item pode assumir mais de duas alternativas, o valor $\sum pq$ é substituído por $\sum s_i^2$, a soma das variâncias de cada item. Esta fórmula genérica é a seguinte:

$$r_{tt} = \left(\frac{n}{n-1}\right)\frac{s_t^2 - \sum s_i^2}{s_t^2}$$

onde, s_t^2 é a variância de todo o teste e

$\sum s_i^2$ o somatório das variâncias de cada item do teste.

Um instrumento submetido à série de análises acima mencionadas pode ser considerado um instrumento válido e fidedigno e pronto para uso na pesquisa. No caso do instrumento ser orientado para uso clínico (casos individuais), ele deve ser submetido à normatização para se poder interpretar os resultados que ele produz. Contudo, para fins de pesquisa, que tipicamente trabalha com comparações de grupos

de sujeitos, esta normatização não é necessária. Aliás, ela não acrescenta nada de novo e útil para a qualidade psicométrica do instrumento; apenas ela é útil para a interpretação dos resultados, pois ela constitui uma simples transformação dos resultados brutos do instrumento em resultados de alguma maneira padronizados.

Bibliografia

Anastasi, A. (1988). *Psychological testing*. 6th ed. New York: Macmillan Pub. Co.

Bendig, A.W. (1954). Reliability and the number of rating scale categories. *Journal of Applied Psychology, 38*, 38-40.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Ford & M. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychological Bulletin, 56*, 81-105. Cap. 6 de D.N. Jackson & S. Messick (1967), *Problems in human assessment*. New York: McGraw-Hill Book Co.

Campbell, D.T. & Stanley, J. (1973). *Experimental and quasi-experimental designs for research*. Skokie, IL: Rand McNally.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cronbach, L.J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.

Goldsamt, M.R. (1971). Effects of scoring method and rating scale length in extreme response style measurement. Unpublished doctoral dissertation, University of Maryland.

Gulliksen, H. (1987). *Theory of mental tests*. Hillsdale, NJ: Erlbaum.

Hambleton, R.K., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of Item Response Theory*. Beverly Hills, CA: SAGE.

Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and applications*. Norwell, MA: Kluwer Nijhoff.

Hambleton, R.K. & Zaal, J.N. (eds. - 1991). *Advances in educational and Psychological testing: Theory and applications*. Boston, MA: Kluwer Academic Publishers.

Harman, H.H. (1967). *Modern factor analysis*. Chicago, IL: University of Chicago Press.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). *Item Response Theory. Applications to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

Jones, R.R. (1968). Differences in response consistency and subject's preferences for three personality inventory response formats. *Proceedings of the 67th Annual Convention of the American Psychological Association, 3*, 247-248.

Komorita, S.S. (1963). Attitude content, intensity, and the neutral point on a Likert scale. *Journal of Social Psychology, 61*, 327-334.

Kuder, G.F. & Richardson, M.W. (1973). The theory of estimation of test reliability. *Psychometrika, 2*, 151-160.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology , 140*, 55 ps.

Lord, F.M. (1980). *Applications of Item Response Theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Mager, R.F. (1981). *Medindo os objetivos de ensino* ou "conseguiu um par adequado". Porto Alegre, RS: Editora Globo.

Matell, M.S. & Jacoby, J. (1972). Is there an optimal number of alternatives for Likert-scale items? *Journal of Applied Psychology, 56(6)*, 506-509.

Matell, M.S. & Jacoby, J. (1971). Is there an optimal number of Likert scale items? Study I: Reliability and validity. *Educational and Psychological Measurement, 31*, 657-674.

Muñiz, J. (1990). *Teoría de respuesta a los ítems*. Madrid: Pirámide.

Muñiz, J. (1992). *Teoría clássica de los tests*. Madrid: Ediciones Pirámide, S.A.

Muñiz, J. (1996). *Psicometría*. Madrid: Univérsitas.

Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement, 22*, 287-293.

Nevo, B. & Sfez, J. (1985). Examinees' feedback questionnaires. *Assessment and Evaluation in Higher Education, 10*, 236-249.

Novick, M.R. & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometricka, 32*, 1-13

Nunnally, J.C., Jr. (1978). *Psychometric theory*. New York: McGraw-Hill.

Pasquali, L. (Org. - 1996). *Teoria e métodos de medida em ciências do comportamento*. Brasília, DF: INEP.

Pasquali, L. (1998). *Psicometria: Teoria e Aplicações*. Brasília, DF: Editora UnB.

Pasquali. L. (1999). Análise fatorial: Um manual teórico-prático. Brasília, DF: Editora UnB.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Dinamarca: Danish Institute for Educational Research.

Rulon, P.J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review, 9*, 99-103.

Santisteban, C. (1990). *Psicometría*. Madrid: Norma.

Thorndike, R.L. (1982). *Applied psychometrics*. Boston, MA: Hougton Mifflin.

Van der Veer, F., Howard, K.I., & Austria, A.M. (1970). Stability and equivalence scores based on three different response formats. *Proceedings of the 78th Annual Convention of the American Psychological Association, 5*, 99-100.

Yela, M. (1987). *Introducción a la teoría de los tests*. Madrid: Facultad de Psicología, Universidad Complutense.

Utilização de Escalas de Razão de Variáveis Clínicas e Sociais

Fátima Aparecida Emm Faleiros Sousa
Ricardo Kamizaki
José Aparecido Da Silva

I. Introdução

Conceitos e fenômenos subjetivos, tais como atitudes sociais, opiniões e processos de julgamentos têm sido difíceis de serem mensurados acuradamente. Muitos conceitos ou variáveis nas Ciências Sociais e da Saúde são de natureza subjetiva e, nessas profissões, são enfrentados muitos problemas para obter medidas de tais variáveis.

A metodologia psicofísica, especialmente os procedimentos de estimação de magnitude e de emparelhamento intermodal, desenvolvidos na psicofísica sensorial e sendo atualmente usado nas Ciências Sociais e da Saúde, tem se mostrado promissora como um instrumento para escalonar fenômenos subjetivos. Nosso propósito, ao rever a literatura sobre o tem, é descrever como técnicas de mensuração, o paradigma teórico sobre o qual elas são baseadas e também vários estudos de âmbito social e clínico, nos quais foram utilizadas essas estratégias de mensuração. Esse enfoque tem sido denominado por alguns autores de Psicofísica Clínica e/ou Social, podendo-se, baseados neste, estabelecer-se o quanto (quantitativamente) um atributo é maior do que o outro e não somente afirmar que são apenas diferentes, qualidade esta atribuída às escalas de categorias (Faleiros Sousa, 1993; Faleiros Sousa & da Silva, 1996).

A revisão de literatura foi dividida em três partes. A primeira, com o objetivo de mostrar a transição e a adaptação da metodologia psicofísica aplicada no domínio sensorial para o social e/ou clínico, abordamos a função de potência ou Lei de Stevens e os métodos psicofísicos de estimação de magnitude e de emparelhamento intermodal. Na segunda, mostramos como esta metodologia pode ser adaptada para mensurar atributos não métricos (sociais e clínicos) e revisamos os experimentos nos quais ela tem sido usada com sucesso. E na última, discutimos vantagens desta metodologia.

1. Acerca da Função de Potência (Lei de Stevens)

No domínio da Psicofísica, um ramo experimental da Psicologia que lida com a mensuração e a análise dos mecanismos e/ou processos subjacentes às

diferentes respostas sensoriais e/ou perceptivas, é bem conhecido que a relação entre as estimativas numéricas (R) e os valores das intensidades físicas dos estímulos (E) é descrita por uma função de potência. Esta função em sua forma mais simples pode ser escrita como:

$$R = k \cdot E^n \quad (1)$$

sendo *k* uma constante arbitrária que depende da unidade de medida empregada e *n* é o expoente da função. O expoente é o parâmetro mais importante, uma vez que determina a curva que representa a relação entre o estímulo e a resposta. Se o expoente é exatamente igual a 1,0, a função segue uma linha reta. Neste caso, a magnitude da sensação registrada (resposta) varia linearmente com a intensidade do estímulo. Quando o expoente é maior do que 1,0, a curva que representa esta função é monotonicamente crescente. Se o expoente é menor do que 1,0, a curva é monotonicamente decrescente. Na Figura 4-1, estas três curvas estão representadas em coordenadas logarítmicas (ver Stevens, 1975; Baird & Noma, 1978; Gescheider, 1997).

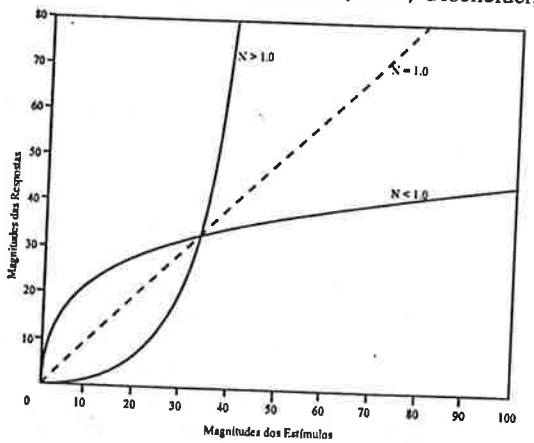


Fig. 4-1. Magnitudes subjetivas em função das magnitudes físicas dos estímulos em coordenadas lineares. A curva é positiva ou negativamente acelerada, dependendo de se o expoente é maior ou menor que 1,0. Quando o expoente é igual a 1,0, a função de potência é uma linha reta em coordenadas lineares. As unidades das escalas têm sido escolhidas arbitrariamente para mostrar a forma relativa das curvas (Stevens, 1975, p. 16)

Colocando-se ambos os termos da Equação 1 em logarítmos, é obtida uma função linear que facilita determinar os parâmetros da função de potência bem como o coeficiente de determinação da função, o qual representa o grau de ajustamento dos dados obtidos. Então,

$$\log R = \log k + n \log E \quad (2)$$

Na Equação 2, o expoente *n* torna-se a inclinação da função linear, enquanto o logaritmo da constante escalar *k* torna-se a intersecção com o eixo das respostas.

Quando a curva é projetada em coordenadas log-log, a relação é representada por uma linha reta, independente do expoente ter um valor maior ou menor que 1,0. Este artifício matemático de projetar em coordenadas logarítmicas faz com que as curvas desapareçam e, por conseqüência, o valor do expoente é refletido diretamente na inclinação da reta. Esta relação está representada na Figura 4-2.

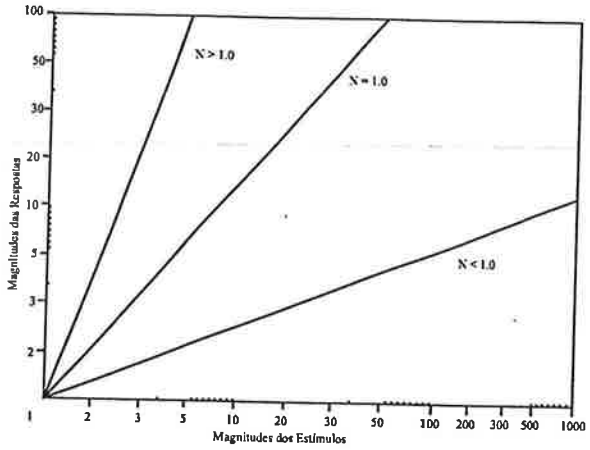


Fig.4- 2. Magnitudes subjetivas em função das magnitudes físicas dos estímulos em coordenadas logarítmicas. As curvas da Figura 4-1 tornam-se linhas retas em coordenadas logarítmicas. A inclinação da linha corresponde ao expoente da função de potência característico de cada modalidade ou atributo sensorial. (Stevens, 1975, p. 17).

A Equação 1 é conhecida como Lei de Potência ou Lei de Stevens e tem sido verificada para uma grande variedade de dimensões aditivas, tais como: sonoridade, brilhância, peso, comprimento, área, volume e distância visual e, nos últimos anos, tem sido estendida para quantificar atributos sociais ou clínicos que por natureza são estímulos não-métricos. Além disso, ela tem sido examinada sob muitas e diferentes condições de estímulos e com ampla variedade de métodos (ver, por exemplo, Da Silva & Macedo, 1982; Engelmann, 1966; Gescheider, 1988; Krueger, 1989; Stevens, 1975). Na tabela 4-1 estão apresentados diferentes valores de expoentes para diferentes modalidades perceptuais e/ou sensoriais escalonadas por métodos baseados em julgamentos de razão.

Diferentes métodos psicofísicos podem ser utilizados para calcular os parâmetros da função de potência (ver, por exemplo, Da Silva & Macedo, 1982; Fukusima, Ribeiro & Da Silva, 1988). Todavia, a validade da escalas sensoriais e/ou perceptivas geradas a partir do método de estimação de magnitude tem sido questionada por vários pesquisadores devido à confiabilidade no uso de números para expressar as respostas sensoriais (Warren & Warren, 1963; Poulton, 1968; Krueger, 1989).

Tabela 4-1. Média e desvio-padrão dos expoentes da função de potência para uma variedade de modalidades sensoriais e/ou perceptivas baseados em julgamentos de razão.

Modalidade ou atributo I	Expoente médio I	Desvio-padrão
Som	0,56*	0,13
Sabor (sacarina)	0,53	0,29
Sabor (salgado)	1,07	0,41
Sabor (amargo)	0,57	0,23
Sabor (azedo)	0,94	0,28
Odor (amil-acetato)	0,19	0,14
Odor (benzeno)	0,56	0,01
Tempo	0,91	0,18
Brilho	0,26	0,13
Temperatura (frio)	1,04	0,05
Temperatura (quente)	1,14	0,37
Numerosidade	0,84	0,22
Número	0,72	0,21
Peso	1,18	0,17
Aceleração angular	1,39	0,52
Choque elétrico	2,80	0,99
Comprimento visual	1,01	0,27
Área visual	0,77	0,16
Volume visual	0,72	0,10
Distância visual	0,97	0,22

* Valor correspondente a 0,28 em termos de intensidade sonora.
Fonte: Da Silva, J. A. & Macedo, L. (1982).

2. O Método de Estimação de Magnitude

O método de estimação de magnitude tem sido freqüentemente usado para escalonar diferentes modalidades perceptivas devido a sua rapidez de aplicação e fácil compreensão por observadores adultos e mesmo por crianças que já têm adquirido o conceito de razão entre objetos ou dimensões. Neste método, o observador recebe previamente instruções e atribui números a uma seqüência de estímulos (métricos ou não métricos) apresentados individualmente, para que faça um julgamento da magnitude percebida de cada um, de forma que esses números reflitam sua impressão subjetiva dos estímulos.

Segundo Stevens (1971), o método de estimação de magnitude denominado por ele de emparelhamento numérico, é também uma forma de emparelhamento intermodal (ver descrição a seguir), sendo que os números são equiparados aos estímulos. Neste contexto, o contínuo de números pode ser considerado como outra modalidade perceptual. Os sujeitos trazem na sua experiência de vida os números consigo, e o experimentador precisa somente fornecer os estímulos com os quais os números serão emparelhados (associados). A natureza da tarefa pode ser retratada num conjunto típico de instruções escritas para o participante.

Uma série de estímulos são apresentados aos participantes numa ordem aleatória para cada um. A tarefa é dizer como esses estímulos parecem assinalando números a eles. Então, sucessivos números proporcionais são assinalados de modo a refletir a impressão subjetiva do participante. Dois são os tipos de estimação de magnitude: com a presença do módulo e com módulo livre. No primeiro tipo, um estímulo é apresentado pelo experimentador como estímulo padrão e a ele é designado um valor numérico denominado módulo ou valor de referência. Em seguida, o observador deve assinalar aos estímulos subsequentes números que sejam proporcionais ao atribuído a esse módulo, os quais representarão a razão julgada entre os diferentes estímulos apresentados pelo experimentador. Dessa forma, se um estímulo da série apresentada é considerado como tendo o dobro da intensidade daquele que recebeu o módulo, ele deve receber um valor numérico que seja duas vezes aquele atribuído ao estímulo padrão (módulo). No segundo tipo, o método de estimação de magnitude não tem um estímulo padrão estabelecido previamente, ou seja, o módulo é livre e o observador poderá assinalar qualquer número ao primeiro estímulo apresentado, devendo os números assinalados para a série de estímulos refletirem razões (ou proporções) entre os estímulos julgados.

O estímulo padrão deve ser escolhido entre os estímulos de média intensidade. Pode ser planejado no experimento, 1 ou 2 julgamentos para cada estímulo, para cada sujeito e ser calculada a média entre eles. Não há necessidade de treinar os sujeitos, uma vez que não há expectativas de respostas certas ou erradas. Sob circunstâncias de dificuldade de alguns sujeitos, pode ser feito treinamento com contínuos mais fáceis, como por exemplo: comprimentos de linhas.

As médias podem ser calculadas através da média geométrica. O expoente determinado pelos logaritmos das médias geométricas não é afetado pelo fato de que cada participante usa uma unidade diferente de módulo.

Em síntese, independente dessas duas variantes do método de estimação de magnitude, representadas pela presença ou ausência do módulo, a razão entre os números assinalados deve refletir a razão entre as intensidades percebidas dos estímulos julgados. Por exemplo, se um estímulo parece 20 vezes mais intenso do que o primeiro julgado (ou estímulo padrão, se for o caso), deve ser assinalado a ele um número 20 vezes maior. Podem ser usados frações, números inteiros, números decimais, mas assinalados proporcionalmente. Todos os demais estímulos devem ser julgados da mesma forma.

3. O Método de Produção de Magnitude

O experimentador apresenta os números, um de cada vez em ordem aleatória para cada sujeito, o qual ajusta os estímulos para produzir um emparelhamento. Os números devem aproximar-se de uma progressão geométrica. Por exemplo, num julgamento de sonoridade, os números sucessivos apresentados foram na razão de 2 para 1 e variaram de 1,25 a 80 (Stevens & Guirao, 1962). Geralmente, devido ao efeito de regressão, a função de potência obtida pela produção de magnitude tem os expoentes maiores do que os obtidos pela estimação de magnitude. Quando os resultados da estimação de magnitude e da produção de magnitude são combinados, trazem mais vantagens do que se calculados isoladamente. Portanto, calculamos a média aritmética dos dois expoentes resultantes para minimizar viéses.

4. Princípios do Emparelhamento Intermodal

O emparelhamento intermodal é considerado o método mais elegante criado por Stevens e colaboradores (Stevens, 1959; J.C. Stevens, Mack & Stevens, 1960) para validar a Lei de Potência e as escalas de magnitude. Sabemos que toda forma de mensuração é um exercício de emparelhamento. O homem primitivo contava o número de cabeças de gado do seu rebanho emparelhando-o com pedras. Uma distância é normalmente medida através do emparelhamento com os múltiplos de uma unidade de medida (no caso, comprimento), seja ela arbitrária ou padronizada, e assim por diante. Emparelha-se a intensidade da sensação com uma unidade qualquer. Assim, submetido a um estímulo, o observador sente a sua intensidade e emparelha-a com aquilo que lhe é mais familiar: o contínuo de número. Esse procedimento possibilita a construção de uma escala de sensação cuja relação com a escala de estímulo pode ser representada por uma Lei Psicofísica, no caso a Lei de Potência ou Lei de Stevens.

Podemos resumir o problema do emparelhamento intermodal na seguinte questão: É possível emparelhar, um ao outro, dois contínuos sensoriais diferentes e independentes, em lugar de emparelhar cada um deles separadamente ao contínuo de números? Para responder a esta questão, Stevens construiu, em 1958, um equipamento que permitia, ao mesmo tempo, estimular independentemente duas diferentes modalidades. Ao observador cabia manipular o controlador de um dos estímulos de modo que a sensação provocada por este parecesse ter magnitude igual à causada por outro. Em outras palavras, este equipamento proporcionava um teste de transitividade da escala sensorial em que,

$$\text{se } A = B \text{ e } A = C, \text{ então } B = C$$

Em notação fundamental, se E_f e E_s são dois contínuos emparelhados a R, o teste tradicional de transitividade pode ser ilustrado para a Lei de Stevens, como se segue:

$$R_f = f(E_f) \text{ e } R_s = g(E_s)$$

então:

$$f(E_f) = g(E_s)$$

Usando esta propriedade de transitividade podemos ilustrá-la tomando dois contínuos sensoriais facilmente manipuláveis. Por exemplo, suponha que um observador faça estimativas de magnitudes de duas modalidades sensoriais e/ou perceptivas diferentes; sejam elas, força dinamométrica e som e, em seguida, funções de potência são ajustadas do modo usual. Desta forma:

$$R_f = E^a \text{ e } R_s = E^b$$

onde (f) e (s) são, respectivamente, as notações para força dinamométrica e som, e (a) e (b) são os expoentes característicos destas duas modalidades (o fato de k ter sido omitido a partir da Equação 1 não afeta o argumento acima). Então, quando R_f e R_s são emparelhados em vários níveis de intensidade a equação resultante será:

$$E^a = E^b$$

e, em seguida, colocando ambos os termos em logaritmos podemos reescrevê-la como:

$$a \log E_f = b \log E_s \text{ ou } \log E_f = (b/a) \log E_s$$

Os expoentes (a) e (b) são empiricamente determinados por procedimentos separados usando o método de estimação de magnitude numérica. Quando os valores de E_s e E_f são plotados em coordenadas log-log a expressão acima representa uma linha reta, ou seja, uma função de potência com a inclinação igual à razão (b/a) dos expoentes originais obtidos através do método de estimação de magnitude. Como os expoentes podem se calculados a partir de dados obtidos de experimentos, esta relação é plenamente testável e pode ser feita em ambas as direções, como no exemplo de som emparelhado à força dinamométrica e vice-versa, para corrigir o efeito de regressão, ou seja, a tendência que os observadores têm de comprimir a variável que está sob seu controle. Quando o expoente obtido aproxima-se daquele predito pela propriedade de transitividade anteriormente descrita, a escala obtida é validada. De fato, Stevens (1975) mostra que a diferença entre ambos dada em decilog é na maioria das vezes menor que 5%.

5. Contínuo Metatético Versus Contínuo Protético

À luz da distinção entre contínuos metatético e protético, Stevens (1974, 1975) também ilustra a atuação dos mecanismos ou processos fisiológicos subjacentes à função de potência. Para Stevens, podemos julgar tanto a qualidade quanto a quantidade de um dado estímulo. Por exemplo, podemos julgar a tonalidade de um som como alta ou baixa, ou seja, sua qualidade; e também julgar a intensidade sonora de um estímulo, isto é, os graus de magnitude ou quantidade. Podemos também julgar o comprimento físico de uma linha e também se a linha está à direita ou à esquerda. Assim, para Stevens quando julgamos a qualidade devemos pensar num contínuo metatético e quando julgamos a quantidade, num contínuo protético.

O critério fundamental que nos permite distinguir estes dois contínuos é como eles se comportam nos experimentos psicofísicos. Stevens nomeou-os em função da natureza dos processos fisiológicos que parecem estar subjacentes a algumas modalidades perceptivas e/ou sensoriais. As intensidades dos estímulos têm sido denominadas de protético porque uma grande quantidades delas é baseada em processos fisiológicos em que a nova excitação é adicionada à excitação já existente. Por exemplo, supõe-se que uma luz parece mais brilhante do que é ou que um som parece mais alto do que é porque uma nova excitação neural é adicionada. De outro lado, um processo de substituição parece estar subjacente aos contínuos denominados metatéticos. Por exemplo, quando a tonalidade parece modificar-se não é por causa de uma excitação que foi adicionada, mas porque uma nova excitação substitui a excitação removida.

A distinção entre estes dois tipos de contínuos é fundamental dentro da teoria psicofísica moderna, pois diferentes leis gerais parecem governar estes dois

diferentes tipos de reações sensoriais. De fato, e como mencionamos nos quatro Estudos, quando as magnitudes subjetivas dos contínuos metatéticos são projetadas em coordenadas logarítmicas em função das intensidades físicas, a relação invariavelmente obedece a Lei de Fechner, e quando as magnitudes subjetivas dos contínuos protéticos são projetadas em coordenadas logarítmicas em função das intensidades físicas a relação via de regra segue a Lei de Stevens. Talvez, as duas leis ou os processos subjacentes a cada uma delas não sejam completamente independentes (Engelman, 1966).

6. O Paradigma do Emparelhamento Intermodal Aplicado aos Atributos Sociais e Clínicos

A questão que agora pode surgir é como este paradigma pode ser aplicado na mensuração de estímulos não métricos, tais como os atributos sociais e os clínicos? Vamos ilustrar hipoteticamente através da seguinte simulação: suponha que observadores sejam instruídos a emparelharem duas modalidades sensoriais quaisquer (por exemplo, força dinamométrica e comprimento de linha) a diferentes atributos sociais não métricos. Isto feito, deve-se plotar os diferentes valores emparelhados (aos atributos não métricos) destas duas dimensões físicas, os de um (força dinamométrica) em função dos do outro (comprimento de linha) e imediatamente estimar os parâmetros da função de função de potência originada da reta de regressão. Estes valores podem ser comparados com aqueles obtidos por Stevens (1975), envolvendo o emparelhamento entre as duas modalidades puramente sensoriais com dimensões físicas mensuráveis ou com aqueles obtidos da reta de regressão dos próprios observadores ou de uma amostra similar, derivados de uma tarefa denominada de calibração.

Em outras palavras, para medir a intensidade de um estímulo social, cada sujeito deve ser instruído a apertar um dinamômetro calibrado, de modo que sua impressão de força dinamométrica seja igual à sua força de impressão do estímulo social; ou seja, quanto mais forte a impressão do estímulo social tanto mais forte deve ser a força dinamométrica. Similarmente, usando-se intensidade do som como modalidade de resposta, o observador deve variar a intensidade do som através de um potenciômetro de um par de fone de ouvidos, de forma que a intensidade percebida seja emparelhada com a impressão do estímulo social, de tal modo que quanto mais forte a impressão do estímulo social tanto maior deve ser a intensidade do som (Lodge, 1982).

Tal como ocorre com estímulos métricos, quando duas ou mais modalidades de respostas de magnitude são emparelhadas a um mesmo conjunto de estímulos sociais, o princípio subjacente a esse relacionamento é o de que intensidades iguais a uma mesma intensidade são iguais uma à outra. Assim, uma escala de magnitude subjetiva é validada pelo método de emparelhamento intermodal quando a inclinação obtida dos emparelhamentos com um conjunto comum de estímulos sociais se aproxima da inclinação obtida a partir da razão entre as duas inclinações características das duas medidas de respostas psicofísicas (Cross, 1974). Devemos ressaltar que a razão predita (teste de critério para validar a escala de magnitude) é uma função entre os dois contínuos (ou modalidades) de respostas e não entre dois estímulos, uma vez que os sujeitos estão usando as duas respostas para expressar suas impressões das intensidades dos estímulos. No escalonamento social, uma primeira alternativa seria a de comparar o expoente empírico (inclinação derivada), quando as modalidades de respostas são emparelhadas com estímulos sociais, com o expoente teórico (inclinação teórica) característico dos relacionamentos dessas modalidades quanto emparelhadas com estímulos físicos. Portanto, para ser validada a escala de atributos sociais, a razão empírica obtida quando os sujeitos fazem o emparelhamento das duas modalidades de respostas com os estímulos sociais, deve ser então a mais próxima da razão estabelecida para essas duas modalidades de respostas emparelhadas a estímulos (Lodge, 1982).

Uma segunda alternativa seria a comparação desse expoente empírico derivado das estimativas dos estímulos sociais com o expoente empírico obtido num experimento de calibração, onde os mesmo sujeitos tenham emparelhado as mesmas duas modalidades de respostas com estímulos métricos. Esse experimento de calibração envolve uma tarefa de escalonamento psicofísico que serve como treino dos observadores no uso das duas modalidades de respostas para que façam julgamentos proporcionais. Essas mesmas duas modalidades deverão ser empregadas num segundo experimento de escalonamento de magnitude, no qual os observadores julgarão a intensidade de um dado estímulo social.

O pressuposto teórico-experimental é que os mesmos viéses que afetam as respostas aos estímulos sensoriais (métricos) atuariam de modo análogo nas respostas aos estímulos sociais (não-métricos). Qualquer que seja a alternativa de comparação, o importante é que cada um dos expoentes empíricos e a razão entre eles precisam ser funções de potência. Sendo assim, quando esse critério é satisfeito a escala derivada é uma escala de razão, dita psicofisicamente validada (Baird & Noma, 1978; Stevens, 1975).

7. Mensuração de Atributos Sociais

Nesta secção apresentaremos sucintamente experimentos nos quais foram escalonados atributos sociais (métricos) através de rigorosos métodos psicofísicos desenvolvidos no domínio da psicofísica sensorial. O objetivo é exemplificar a riqueza da diversidade metodológica utilizada para o escalonamento de contínuos dessa natureza, bem como descrever o processo de validação psicofísica envolvido nesses escalonamentos.

Importância política dos monarcas suecos

Em estudo feito por Ekman e Künnapas (1963) foram utilizadas estimativas de razão e de comparação aos pares para construir escalas da importância política de 11 monarcas suecos, com os quais os observadores estavam familiarizados. Os resultados comparativos demonstraram que quando os valores estimados do escalonamento de comparação aos pares foram representados graficamente em coordenadas lineares em função dos valores estimados do escalonamento de razão, o gráfico foi curvilíneo. Posteriormente, quando os valores escalares derivados do método de comparação aos pares foram representados graficamente em função dos logarítmos das estimativas de razão, o resultado foi uma linha reta. Portanto, este padrão de resultados é similar àquele usualmente obtido quando ambos os procedi-

mentos psicofísicos são empregados para escalonar dimensões métricas. Tais resultados mostram que o contínuo não métrico da importância política dos monarcas suecos é um contínuo quantitativo e não qualitativo.

Valores estéticos de desenhos e de manuscritos

Dois estudos sobre os valores estéticos de manuscritos foram realizados por Ekman e Künnapas (1960, 1962a). Em ambos, amostras de manuscritos foram escalonadas pelo método que comparação aos pares e de estimação de razão. Tomados juntos os dados mostraram que também para este tipo de contínuo, tal como ocorre com alguns contínuos métricos, a amplitude dos estímulos é um fator importante em determinar a relação funcional entre os valores escalares derivados de métodos diferentes. Embora tenham mostrado que uma relação logarítmica descreve muito bem a relação entre os valores escalares derivados do método de comparação aos pares e os valores escalares derivados do método de estimação de razão, a forma da função é frequentemente mascarada quando a amplitude é pequena.

Num estudo relacionado, Ekman e Künnapas (1962b) empregaram os métodos de estimação de categorias, de razão e de comparação aos pares para escalonar desenhos de diferentes árvores. Os resultados novamente mostraram que os valores escalares originados dos métodos de estimação de categorias e de comparação aos pares são logaritmicamente relacionados aos valores escalares derivados do método de estimação de razão. Todavia, os valores escalares obtidos nos dois primeiros métodos são linearmente relacionados entre si. Portanto, também para este contínuo a mesma relação usualmente encontrada com dimensões métricas é fortemente estabelecida.

Preferências musicais

Julgamentos estéticos na esfera musical foram investigados por Kohn (1965), o qual utilizou os métodos de estimação de magnitudes e de estimação de categorias. Diferentes grupos de sujeitos julgaram seleções vocais e peças de piano. Os resultados mostraram que em coordenadas mono-log, as correlações produto-momento entre os valores escalares das estimativas de magnitude variaram de 0,90 a 0,96. Embora tenha havido uma concavidade ascendente que usualmente ocorre quando as escalas de categorias são projetadas em função dos logaritmos das escalas de magnitude, a grande variabilidade dos valores impediu que a mesma fosse mais saliente. Os dados também mostraram que esta relação entre as escalas de categorias e escalas de magnitudes foi invariante a despeito de diferenças na idade, sexo, educação, ocupação e patologia dos sujeitos. Devido a isso, Kohn (1965) destacou que essas invariâncias empíricas sugerem a utilidade e a robustez do método de estimação de magnitude em revelar processos de julgamentos complexos.

Preferências por relógios de pulso

Indow (1961) apresentou figuras e descrições de vários relógios de pulsos para estudantes universitários japoneses e solicitou-lhes que fizessem julga-

mentos pelo método de comparação aos pares e estimativas de razão. Na realidade, os estudantes emparelharam o comprimento de uma linha ao valor subjetivo de sua preferência, ou seja, realizaram um tipo de emparelhamento intermodal. Comparando os valores escalares obtidos por meio do comprimento de linha com os valores escalares derivados do método de comparação aos pares, Indow demonstrou uma relação aproximadamente logarítmica entre os dois tipos de valores escalares.

Numa outra parte do experimento, Indow (1961) solicitou aos estudantes para indicar, em yens, qual seria o preço justo para cada um dos relógios de pulso. Os resultados mostraram que a a relação entre as estimativas médias dos preços e os valores escalares de preferência, em coordenadas log-log, segue uma função de potência com um expoente igual a 0,32. Este valor indica, portanto, que a relação entre o preço estimado e a preferência não é linear.

Julgamento moral

Ekman (1962) usou os métodos de comparação aos pares e o de estimação de razão para escalonar diferentes atitudes, descritas verbalmente, representando comportamentos mais ou menos imorais ou criminosos. Os resultados, semelhantes àqueles obtidos em outros estudos, mostraram uma relação logarítmica entre os valores escalares derivados dos julgamentos obtidos pelo método de comparação aos pares e os dos julgamentos derivados das estimativas de razões. Portanto, o contínuo de julgamento moral também pode ser considerado como quantitativo, se considerarmos como critério a obtenção de uma função logarítmica entre os dados dos diferentes métodos (mensuração de diferenças e de razões).

Comportamento racista

Dawson e Brinker (1971) fizeram um estudo em que solicitaram aos sujeitos que julgassem o quão racista eram sentenças que expressavam diferentes comportamentos de uma pessoa branca em relação à negra. A tarefa dos sujeitos consistia em indicar sua opinião de quão racista era o comportamento descrito através do método de emparelhamento intermodal envolvendo as modalidades de som e força dinamométrica. Para analisar os resultados, os valores dos emparelhamentos de som foram colocados em função dos valores dos emparelhamentos de força dinamométricas, em coordenadas logarítmicas. Tais dados mostraram claramente uma função de potência com um expoente igual a 0,39, indicando que a pressão sonora é uma função de potência da força dinamométrica. Este valor obtido experimentalmente é muito próximo daquele de 0,38, que é o valor esperado a partir da razão entre os valores de 0,64 (expoente para som) e 1,70 (expoente para força dinamométrica). Portanto, estes resultados sugerem que os sujeitos podem indicar os graus de suas opiniões emparelhando-as a duas intensidades sensoriais diferentes. Eles também são consistentes com aqueles obtidos quando emparelhamentos envolvendo estas duas modalidades foram feitos diretamente, uma com a outra (ver também, Stevens, 1966).

Grau de liberalismo-conservadorismo

O julgamento do grau de liberalismo-conservadorismo expresso em diferentes afirmações é um contínuo cuja relação entre as estimativas de categorias (ou de comparação aos pares) e estimativas de razões não segue uma função aproximadamente logarítmica. Este continuo foi investigado por Ekman e Künnapas (1963), os quais realizaram dois experimentos envolvendo diferentes números de afirmações. Os dados de ambos os experimentos mostraram que a relação entre os valores escalares obtidos pelo método de estimação de razões é claramente linear em função dos valores escalares obtidos pelo método de comparação aos pares. Uma interpretação da diferença entre estes resultados com aqueles obtidos com outros contínuos, proposta por Ekman e Künnapas, é que o grau de conservadorismo representa uma posição sobre um contínuo qualitativo que se estende de um liberalismo típico a um conservadorismo típico, enquanto os contínuos anteriores representam a quantidade de beleza, imoralidade e influência política. Portanto, nesse sentido o contínuo de conservadorismo pode ser qualitativo e os contínuos de valores estéticos, julgamento moral e prestígio político podem ser considerados quantitativos, se focalizarmos a relação funcional entre as estimativas de diferenças (estimativas de categorias ou comparação aos pares) e as estimativas de razões (estimativas de magnitudes).

Percepção do poder nacional

Um estudo dos mais interessantes envolvendo escalonamento social foi realizado por Shinn (1969). Ele quantificou as opiniões a respeito do poder nacional de várias nações. Cada nação foi descrita em termos de três atributos: população, produto nacional bruto e porcentagem do PNB destinado à militarização. Os resultados mostraram que todos os três atributos contribuíram para o poder nacional aparente, mas o crescimento do poder nacional cresceu de modo diferente para quantidades crescentes de cada um dos três atributos. Foi demostrado que os dados podem ser descritos muito bem por três funções de potência, uma diferente da outra. Shinn também verificou como as três variáveis de poder nacional se combinavam para criar o poder nacional total percebido. Para isso foram testados dois modelos, o aditivo e o multiplicativo. O modelo multiplicativo provou ser superior, indicando que a percepção do poder nacional pode ser regida por uma equação multiplicando os três atributos e na qual o poder cresce mais rapidamente quando a produtividade aumenta.

Utilidade subjetiva de bens e de benefícios públicos

Kemp (1988, 1991) empregou o método de estimação de magnitude para investigar a utilidade subjetiva de bens e de benefícios públicos pessoal e nacionalmente consumidos. As medianas das estimativas de magnitudes foram calculadas para cada um dos itens pessoais e nacionais e, em seguida, foram projetadas em função dos custos reais avaliados de cada item. Os expoentes das funções de potência variaram entre 0,13 e 0,46. Além disso, os dados mostraram que os bens pessoais são mais relacionados aos seus respectivos custos do que os benefícios públicos. Os

dados também sugerem que a relação utilidade-custo para benefícios públicos não é a mesma que aquela para bens pessoais. A primeira foi percebida como sendo de utilidade mais alta e a função de potência foi caracterizada por um expoente menor para esses benefícios públicos.

Opiniões sócio-políticas

Lodge e colaboradores (Lodge & Tursky, 1982; Lodge, 1982; Lodge, Cross, Tursky & Tanenhaus, 1975; Lodge, Cross, Tursky, Tanenhaus & Reeder, 1976) realizaram vários experimentos em que investigaram quais atributos as pessoas associam com partidos políticos, candidatos, instituições e polícia, e também quão intensamente as pessoas concebem sobre estes atributos. Por exemplo, quão intensamente uma pessoa se identifica com um partido político? Quanto de confidência uma pessoa tem da Suprema Corte? Quão comprometidas são as pessoas com os processos democráticos?

Para responder a estas questões foi composta uma escala com 30 adjetivos variando desde absolutamente perfeito até desgostoso, os quais constituíram os itens usados para mensurar o suporte político. Diferentes grupos de sujeitos usaram as modalidades de respostas de estimação numérica de forças dinamométricas e de pressões sonoras para avaliar a quantidade de suporte inserido nestes adjetivos. As médias geométricas foram calculadas para cada uma das modalidades de respostas. As correlações entre os logaritmos destas médias foram altas, indicando um alto grau de dependência linear entre as medidas, e cada um dos expoentes empíricos aproximou-se do expoente esperado dentro dos limites de 95% de confiança. Este padrão de resultados constitui uma alta validade de constructo e também forte evidência de um contínuo psicológico subjacente ao suporte político.

Seriedade de ofensas e de crimes

Em 1964, Sellin e Wolfgang publicaram no livro "Mesuração da Delinqüência" os resultados de três anos ininterruptos de investigações usando métodos psicofísicos para mensurar a criminalidade em geral, e a delinqüência, em particular. O delineamento enfatizou eventos delinqüentes e não as pessoas delinqüentes. O propósito principal foi mensurar a quantidade e o tipo de prejuízo para a comunidade gerado por um ato anti-social. Os procedimentos psicofísicos de estimação de categorias e de estimação de magnitudes foram aplicados aos eventos selecionados com o objetivo de converter a seriedade julgada destes eventos em escores numéricos, Tal como ocorreu com outros contínuos não métricos, a relação entre as estimativas de categorias e as estimativas de magnitudes do grau de delinqüência é uma função logarítmica. Quando as estimativas de categorias são projetadas em função dos logarítmos das estimativas de magnitudes a relação é aproximadamente linear, embora com uma leve concavidade ascendente. Tais relações mostram que a seriedade das ofensas é um contínuo quantitativo. Em adição, os resultados de Sellin e Wolfgang (1964) revelaram que tais relações são invariantes tanto em função da idade quanto das amostras dos sujeitos. Portanto, estes resultados indicam que a própria ofensa é aparentemente o determinante principal do julgamento de sua seriedade.

Preferência e prestígio ocupacionais e profissionais

Um dos primeiros estudos que investigou através de métodos psicofísicos escalares o prestígio ocupacional e profissional foi realizado por Perloe (1963). Ele utilizou e comparou os métodos de estimação de magnitudes e estimação de categorias para escalonarem uma lista de 100 ocupações. Os dados revelaram que a escala de categorias de prestígio ocupacional é uma função aproximadamente logarítmica da escala de estimação de magnitudes, quando os sujeitos não limitaram a amplitude de julgamento das profissões consideradas de mais alto prestígio. Portanto, estes dados indicam que o contínuo de prestígio profissional tem características quantitativas. Tais resultados foram posteriormente confirmados por Künnapas e Wikstroem (1963), Dawson e Brinker (1971), Dawson e Mirando (1976) e Hardin e Birnbaum (1990), os quais mostraram também que o expoente empírico derivado da função de potência relacionando o emparelhamento de intensidade de sons a forças dinamométricas não foi diferente daquele predito pela propriedade de transitividade do emparelhamento intermodal. Tomados em conjunto, os dados destes experimentos mostraram que o contínuo de prestígio e de preferências ocupacionais e profissionais possui características quantitativas e que entre estudantes americanos e suecos, a profissão de médico é a que possui o mais alto prestígio ou é a profissão mais preferida.

Na tentativa de introduzir o paradigma da Psicofísica na pesquisa em Enfermagem, Faleiros Souza (1993) mensurou o prestígio profissional do enfermeiro, através de escalonamentos de razão e relatou procedimentos não utilizados ainda nessa área no Brasil. Tomados em conjunto, os dados destes experimentos mostraram que o contínuo de prestígio profissional possui características quantitativas e que as respostas de razão derivadas do método de estimação de magnitude foram substancialmente diferentes daquelas obtidas com escalonamento a nível ordinal quanto às intensidades dos estímulos, portanto, sugerindo que a estimação de magnitudes é superior em detectar variações com níveis de estímulos elevados, uma vez que com este método a amplitude de respostas é ilimitada.

Status social

Diferente dos estudos anteriores nos quais inúmeras ocupações foram mensuradas em termos de seus respectivos prestígios, preferências e desejabilidade, Hamblin e colaboradores (Hamblin & Smith,1966; Hamblin, 1971) investigaram o status local e o profissional de professores universitários. O status local do professor foi avaliado em seu próprio departamento de trabalho, enquanto o status profissional foi avaliado considerando-se a profissão e a disciplina ministrada em relação a similares em outras partes do país. Os resultados mostraram que o status local é uma função de potência multivariada das seguintes variáveis: mérito de ensinar e da liderança do professor. Somente estas duas variáveis explicam 97% da variância dos julgamentos do status local do professor. De outro lado, o status profissional foi uma função de potência multivariada de 4 variáveis independentes: mérito de publicação e mérito de ensinar como as mais relevantes, e cordialidade (negativamente relacionada) e tempo de serviço como as menos relevantes, as quais explicam 99% da variância dos julgamentos do status profissional do professor.

8. Mensuração de Atributos Clínicos

Nesta secção apresentaremos alguns experimentos nos quais foram escalonados atributos fisiológicos e patológicos através dos procedimentos psicofísicos desenvolvidos no domínio da Psicofísica Sensorial. Importante mencionar que estes atributos podem ter ou não dimensões físicas mensuráveis. Por exemplo, de um lado, na determinação da acuidade os estímulos são intensidades de sons mensuráveis fisicamente. De outro lado, na mensuração do stress e/ou reajustamentos sociais, os estímulos são diferentes eventos de vida que não podem ser mensurados fisicamente. O objetivo desta secção é mostrar que o conhecimento metodológico e teórico obtido e desenvolvido no domínio da Psicofísica Sensorial pode ser aplicado em diagnóstico clínico e também na avaliação de fatores, estímulos ou eventos de vida que são estressantes.

Stress e/ou reajustamentos sociais

Holmes e colaboradores (Rahe, Smith, Kjaer & Holmes, 1964; Masuda & Holmes, 1967; Holmes & Rahe, 1967 Ruch & Holmes, 1971) elaboraram uma escala contendo 43 itens de eventos de vida que requeriam diferentes graus de reajustamentos. Alguns desses itens foram: casamento. morte da esposa, gravidez, mudança de residência, dificuldades sexuais, divórcio, problemas no emprego e férias. Esta escala foi denominada por Holmes e Rahe (1967) de escala de magnitude de reajustamentos sociais. A tarefa dos sujeitos consistiu em dar uma estimativa de magnitude para cada um dos itens dessa escala que refletisse o grau relativo de reajustamento necessário para se acomodar a este evento, independente da desejabilidade do mesmo. O evento de vida, casamento, foi tomado como estímulo padrão e a ele foi designado o valor de 500. Os outros eventos deveriam ser estimados proporcionalmente ao casamento tomado como padrão. Os resultados foram extremamente consistentes e invariantes entre diferentes culturas e subculturas de amostras de sujeitos. Posteriormente, a escala foi amplamente utilizada em diferentes contextos, como o hospitalar (Volicer & Bohannon, 1975) e com diferentes faixas etárias, como na velhice (Muheenkamp, Gress & Flood, 1975).

Dois estudos realizados por Holmes ilustram claramente a metodologia psicofísica empregada e permitem afirmar que o contínuo de stress e/ou reajustamento social é um contínuo com características qualitativas ou na linguagem de Stevens, metatético. No primeiro estudo, Masuda e Holmes (1967) solicitaram aos sujeitos para darem estimativas de magnitudes numéricas aos diferentes eventos de vida, tomando como padrão o evento casamento. Os resultados mostraram um coeficiente de concordância muito alto entre as três medidas de tendência central utilizadas: média geométrica, média aritmética e mediana, e também uma relação linear entre o erro padrão e a média geométrica das estimativas de cada item. No segundo estudo, Ruch e Homes (1971) compararam as estimativas de magnitudes com as estimativas de comparações aos pares de 11 itens ou eventos de vida previamente selecionados. Comparação entre os métodos revelou, de um lado, que os valores escalares resultantes das estimativas de magnitudes são altamente correlacionados com os valores escalares derivados das estimativas de comparações aos pares. O coeficiente

de correlação de ordem foi bastante elevado. De outro lado, a relação entre as médias geométricas das estimativas de magnitudes e os valores escalares ajustados resultantes das comparações aos pares foi linear, indicando, portanto, que o contínuo de stress e/ou reajustamento social possui apenas características qualitativas. Posteriormente, Birnbaum e Sotoodeh (1991) e Birnbaum (1992; ver também Crandall, 1992) analisando a mensuração do stress sob o ponto de vista da teoria geral da mensuração, confirmaram que este contínuo é qualitativo e não quantitativo.

Gravidade de enfermidades

O método de estimação de magnitudes também foi utilizado com sucesso na mensuração da gravidade de diferentes enfermidades. De fato, Wyler, Masuda e Homes (1968) elaboraram uma lista contendo 126 enfermidades a qual foi enviada pelo correio a duas amostras distintas: uma não médica e a outra médica. A tarefa dos sujeitos consistia em estimar a magnitude da gravidade das enfermidades assinalando a cada uma delas um número que fosse proporcional ao valor de 500 designado à enfermidade da úlcera péptica. Exemplos de algumas enfermidades foram: constipação intestinal, enxaqueca, diarréia, sinusite, acne, astigmatismo, menopausa, menstruação, eczema, alergia medicamentosa, gonorréia, coma, depressão, epilepsia, derrame cerebral, ataque cardíaco, uremia, câncer e leucemia. As duas amostras foram altamente concordantes em suas estimativas de magnitudes de cada uma dessas enfermidades bem como em suas respectivas ordenações. Os resultados também indicaram que as variáveis idade, sexo, estado civil, etc., afetam numa extensão maior os julgamentos feitos pela amostra não médica do que aqueles feitos pela amostra médica. Combinando as estimativas de magnitudes de ambas as amostras, estas indicaram que a caspa foi a enfermidade com menor estimativa de magnitude, o aborto foi uma enfermidade com estimativa de magnitude mediana e a leucemia foi a enfermidade com maior estimativa de magnitude. O mesmo padrão de resultados foi obtido por Wyler, Masuda e Holmes (1971) e por Volicer e Bohannon (1975).

Em outro estudo similar, Wyler, Masuda e Holmes (1970) replicaram o trabalho original usando uma outra amostra de médicos e analisando as estimativas em função de suas respectivas especialidades médicas. Os resultados mostraram que as estimativas de magnitudes numéricas feitas por médicos de diferentes especialidades não foram significativamente diferentes entre si, exceto apenas para 5 enfermidades, indicando, portanto, que a variável especialidade do respondente não é significativa.

O objetivo principal do estudo realizado por Kamizaki (1997) foi escalonar a severidade de quadros clínicos. Esta escala foi determinada utilizando métodos psicofísicos diretos e indiretos, tais como escalas de razão, escalas intervalares e escalas de ordenações, sendo, posteriormente comparadas. Os objetivos secundários foram verificar se o contínuo não métrico de severidade de quadros clínicos possui características protéticas ou metatéticas, além de verificar se a lei de Ekman é válida para este contínuo não métrico. Três experimentos foram realizados, sendo que o Experimento 1 consistiu da replicação do estudo de Wyler, Masuda e Holmes (1968), no qual 100 quadros clínicos, tais como, verruga, psisoríase, câncer, leucemia e AIDS foram selecionados e apresentados a 47 participantes (20 médicos, 20 enfermeiros e 7 psicólogos). O coeficiente de correlação de Pearson entre as amostras brasileiras variou entre 0,92 e 0,94. No Experimento 2, 15 dos 100 diagnósticos do Experimento 1 foram selecionados e avaliados pelos métodos de estimação de magnitude e de categoria. Os instrumentos foram aplicados a 46 participantes (20 enfermeiros, 16 psicólogos e 10 médicos). Os resultados indicaram que a escala psicofísica de gravidades de quadros clínicos possui características de contínuo protético, além da confirmação da validade da lei de Ekman também para contínuos não métricos. No experimento 3 foram utilizados os métodos de estimações de magnitude e emparelhamento intermodal utilizando-se a modalidade comprimento de linhas. Estes métodos foram aplicados a 31 participantes (10 enfermeiros, 13 psicólogos e 8 médicos). O expoente encontrado foi de 0,93, sendo este valor próximo ao expoente predito numa tarefa de calibração, ou seja 0,99. Em síntese, os dados mostraram uma escala de razão de severidade de quadros clínicos, válida, estável e consistente. Em estudo feito por Faleiros Sousa (1997), foi mensurado a gravidade de quadros clínicos resultantes de cirurgias, através de escalonamento de razão. Tomados em conjunto, os dados mostraram que o contínuo de gravidade de quadros clínicos-cirúrgicos possui características protéticas e que as respostas de razão derivadas do método de estimação de magnitude foram substancialmente diferentes daquelas obtidas através de estimação de categorias, portanto, sugerindo que a estimação de magnitudes é superior em detectar variações com níveis de estímulos elevados, uma vez que com este método a amplitude de respostas é ilimitada.

Deficiência de fala e de pronúncia

Dawson e Brinker (1971) investigaram pelo método de emparelhamento intermodal (som, força dinamométrica e duração temporal) a facilidade de pronunciar diferentes trigramas. Os resultados mostraram que o expoente obtido foi muito similar ao esperado derivado da razão entre os expoentes usualmente obtidos com estimativas de magnitude de som e com força dinamométrica. Também uma alta correlação de ordem indicou que os sujeitos são consistentes em seus emparelhamentos feitos por meio de dois contínuos físicos diferentes. Em outro estudo similar, Dawson e Mirando (1973) verificaram que estas estimativas são estáveis e repetíveis e que a relação entre as estimativas de categorias e as estimativas de emparelhamentos de forças dinamométricas é uma função negativamente acelerada. Em adição, foi observado que a relação entre as estimativas da facilidade e as estimativas da dificuldade de pronunciar diferentes trigramas é uma função de potência com um expoente igual -1,0.

O julgamento do sotaque explícito na pronúncia de uma frase foi investigado por Brennan, Ryan e Dawson (1975) usando entonações diferentes de uma só frase obtidas de diferentes estudantes bilingües (inglês-espanhol). Os resultados mostraram que os expoentes obtidos experimentalmente a partir da função de potência ajustada entre as estimativas de magnitudes e os emparelhamentos de forças dinamométricas, em coordenadas logarítmicas, foram muito próximos do valor esperado a partir da razão entre os expoentes usualmente encontrados para julgamentos de números e para julgamentos de forças dinamométricas.

Mais recentemente, Fucci e colaboradores (Fucci, Ellis & Petrosino, 1990; Ellis & Fucci, 1991) utilizaram o método de estimação de magnitudes para

investigar a estabilidade das escalas obtidas dos julgamentos de clareza e inteligibilidade da fala feitos por fonoaudiólogos e sujeitos inexperientes. Os resultados não indicaram diferenças significativas nas respostas para sentenças sem sentido e com sentido e, além disso, um alto coeficiente de fidedignidade teste-reteste foi obtido entre duas tentativas obtidas numa mesma sessão. Frente a este padrão de resultados, os autores discutem a importância, a potencialidade e as aplicações do procedimento de estimação de magnitudes em pesquisas clínicas e sociais.

Deficiências auditivas

A aplicação de procedimentos psicofísicos a problemas clínicos tem sido extremamente valiosa por fornecer métodos para sistematizar e localizar funções anômalas e para determinar a eficácia de um tratamento. De fato, os métodos psicofísicos tem sido centrais para a análise audiométrica da perda auditiva, para o desenvolvimento de próteses auditivas e para os esforços em usar a visão e o tato como substitutos da audição. Vamos exemplificar como um procedimento psicofísico pode ser utilizado no diagnóstico de deficiências auditivas em pacientes com perda neural e condutiva. De uma amostra de pacientes de uma clínica, Thalmann (1965) selecionou um grupo de 10 pacientes, os quais tinham audição normal em um dos ouvidos e uma grande perda auditiva, aproximadamente igual a 50 dB., no outro ouvido. Cinco pacientes tinham perda condutiva e cinco tinham perda neural. O experimento requeria que os pacientes ajustassem a amplitude de uma vibração aplicada ao dedo indicador para emparelhá-lo à intensidade aparente de um som de 1.000 Hz aplicado a um ou ao outro ouvido do paciente. As médias geométricas dos emparelhamentos foram calculadas e projetadas uma em função da outra. As relação obtidas foram funções de potências, com uma função característica para cada tipo de deficiência auditiva.

Ao analisar estes resultados, Stevens (1975) concluiu que mesmo embora um orgão sensorial possa ser deficiente, um experimento psicofísico bem delineado e conduzido empregando emparelhamentos intermodais pode ser também útil em revelar a natureza da deficiência. Além disso, os métodos psicofísicos tais como estimação de magnitudes e emparelhamento intermodal são muito mais fáceis de serem aplicados do que as tarefas psicoacústicas frequentemente utilizadas em clínicas otológicas. Eles requerem menos explicações, preparos e treinos. Devido, talvez, a estas facilidades de uso, alguns pesquisadores têm empregado métodos psicofísicos com pacientes surdos com o propósito de explorar sua utilidade na seleção de aparelhos de correção auditiva. Por exemplo, Geller e Margolis (1984) e Knight e Margolis (1984) compararam os procedimentos otológicos clássicos de mensuração indireta dos níveis de conforto e desconforto da sonoridade, com o procedimento psicofísico direto de estimação de magnitudes. A grosso modo, os resultados mostraram a vulnerabilidade dos procedimentos otológicos clássicos em função de mudanças nas instruções, procedimento psicométrico e tipo de amplitude de estímulos, indicando que estas medidas podem não ser úteis na seleção de aparelhos auditivos. Esta instabilidade pode ser causada devido às diferentes interpretações de conforto dentre e entre pacientes e clínicos. Ao contrário, os métodos escalares diretos produzem medidas estáveis e repetíveis mesmo entre sujeitos inexperientes e produzem resul-

tados similares entre grupos de sujeitos muito diferentes. Além disso, uma outra aplicação dos métodos psicofísicos é a determinação da sonoridade da fala experienciada por sujeitos normais. Com esta informação, uma estratégia diferente pode ser atingida na seleção de aparelhos auditivos. Isto é, ao invés de determinar os níveis de saída de sons que caem dentro de uma amplitude confortável, torna-se possível determinar a resposta de ganho de freqüência e características de compressão que restauram a sonoridade normal da fala para pessoas surdas (Schiavetti, Metz, & Sitler, 1981). Estes autores mostraram através da comparação entre os métodos de estimação de magnitudes e de estimação de categorias tanto para pessoas com audição normal, quanto para pessoas com surdez congênita que, a inteligibilidade da fala é um contínuo quantitativo. Isto porque para ambos os grupos de sujeitos, a relação entre as estimativas de categorias e as estimativas de magnitudes foi curvilínea em coordenadas lineares.

Dispnéia

A dispnéia definida como a percepção desconfortável e desagradável de dificuldade de respirar, é um fenômeno clínico complexo difícil de ser mensurado. De fato, a dispnéia inclui tanto a percepção de dificuldade de respirar quanto a reação a esta percepção. A dispnéia tem tanto componentes afetivos e cognitivos quanto componentes neurosensoriais e mecânicos. Esses componentes coletivamente contribuem para a percepção da capacidade respiratória (Mahler, Rosiello, Harver, Lentine, McGovern & Daubenspeck, 1987; Harver, 1987; Harver, Tenney & Baird, 1986; Harver & Kotses, 1987; Da Silva & Fukusima, 1989). Cada uma dessas percepções faz parte do constructo de dispnéia em sua totalidade. Não obstante, escalas clínicas têm sido freqüentemente utilizadas para avaliar esta desagradável sensação e recentemente algumas delas têm sido comparadas com métodos psicofísicos diretos tais como estimação de magnitudes tanto com amplitude limitada quanto com amplitude ilimitada (Nield, Kim & Patel, 1989; Nield & Kim, 1991).

Killian, Mahutte e Campbell (1981) usaram a força de resistência externa e a resistência elástica para respiração para determinar se o expoente da função de potência poderia ser confiavelmente estimado. A tarefa dos sujeitos consistiu em designar uma estimativa de magnitude para cada uma das forças de resistência externa e elástica. Os dados revelaram expoentes diferentes para a força de resistência externa e para a capacidade respiratória. Todavia, a correlação produto-momento entre os expoentes individuais obtidos das duas variáveis foi alto, sugerindo que tanto o volume quanto o fluxo de entrada de ar contribuem para a percepção respiratória que ocorre em conexão com a capacidade respiratória. Também, Gottfried, Redline e Altose (1985) exploraram a percepção tanto da força de resistência externa quanto da resistência elástica para a respiração em sujeitos com doença pulmonar obstrutiva crônica e em sujeitos normais, para determinar qual diferença na percepção da capacidade respiratória ocorre quando a magnitude da capacidade é expressa em termos do nível e duração da força muscular respiratória. Os resultados indicando que não houve diferenças entre os dois grupos na percepção de volume inspirado e na força respiratória muscular sugerem que o volume e a força não explicam a redução no expoente. Ao contrário, diferenças em integrar e processar estímulos aferentes no

sistema nervoso central foram as explicações mais prováveis (Nield, Kim & Patel, 1989).

Com estimativas de magnitudes de amplitude limitada, Burdon, Juniper, Killian, Hargreave e Campbell (1982) usaram uma modificação da escala de categoria-razão (estimativas de magnitudes limitadas) de Borg (1982; Borg & Ottonson, 1986) para estabelecer a relação entre a intensidade do esforço respiratório e o grau de obstrução do fluxo aéreo como mensurado pelo volume respiratório forçado em 1 segundo. Os dados obtidos confirmaram que a percepção da intensidade do esforço respiratório aumentou enquanto o volume expiratório diminuiu. Outros estudos têm empregado a escala de categoria-razão como um contínuo de resposta, para quantificar a percepção de intensidade da dispnéia. Em geral, as relações obtidas suportam a teoria que afirma que a percepção da capacidade respiratória é a percepção do esforço muscular respiratório, porque cada variável dependente poderia ser um indicador válido do esforço respiratório (Le Blanc, Bowie, Summers, Jones & Killian, 1986). Também, o método de estimação de magnitudes tem sido empregado para estudar a relação entre fadiga muscular e dispnéia. Os dados mostraram, todavia, que nenhuma relação existe entre a severidade da percepção do esforço respiratório (um indicador da dispnéia) e a presença de um padrão de fadiga diafragmático.

Ansiedade

Wolpe (1969) afirma que a construção de hierarquias, ou seja, de escalas de unidades subjetivas de desconforto, é central à condução sistemática da desensibização para o paciente. Tryon (1977) utilizando essa concepção, investigou a construção de hierarquias para graus de ansiedade. O último autor utilizou o método de estimativas de magnitudes e as escalas de unidades subjetivas de desconforto (suds), nas quais o zero representa a mais absoluta calma e 100, o desconforto máximo, para que os participantes avaliassem o grau de ansiedade decorrentes da variação do número de baratas que encontrassem. O julgamento do sujeito deveria ser feito pressupondo os encontros com quantidades de baratas que variavam de 1 a 24, na própria residência, tanto durante o dia, quanto à noite. Verificou-se pelas estimativas de magnitudes e o suds que até 16 baratas (estímulos), eliciam julgamentos semelhantes. Após este estímulo, os resultados são significativamente diferentes. No estímulo 24 , a estimativa de magnitude foi o dobro da estimativa do suds, mostrando que, mesmo utilizando instrumentos de larga escala (o suds varia de 0 a 100 pontos), os participantes podem comprimir seus julgamentos, e portanto não refletir nesta escala o seu nível de ansiedade.

Fobias

Sullivan (1969, 1970, 1971, 1973) apresentou uma série de experimentos utilizando procedimentos de estimação de magnitude para examinar ansiedades e fobias à cobras. No primeiro estudo, em 1969, o autor convidou 20 estudantes universitários, os quais estimaram um nível de intensidade sonora que Sullivan denominou limiar mínimo aversivo (MAT). Os mesmos sujeitos também estimaram uma escala de ansiedade de 7 pontos que variava de, perfeitamente relaxado, até intensamente desconfortável, e foram argüidos também, quanto ao nível da escala que seria equivalente ao MAT. Os sujeitos emparelharam níveis de intensidades sonoras aos diferentes graus de ansiedades. Os resultados apontaram para a consistência e a fidedignidade dos emparelhamentos entre os dois estímulos ansiogênicos apresentados e a amplitude da intensidade sonora, sugerindo que o emparelhamento intermodal pode medir outros atributos subjetivos no segundo estudo, Sullivan (1970) utilizou um procedimento similar convidando 26 sujeitos para comparar graus de ansiedade durante provas de meio do ano e exame final, utilizando estimação de magnitude e estimação de categoria. O gráfico entre essas estimações apresentou uma concavidade voltada para baixo, indicando que o contínuo possui características protéticas. Este resultado demonstra que a ansiedade, estímulo não métrico, pode ser considerada como um contínuo mensurável pelos métodos psicofísicos. Em 1971 em outro estudo, Sullivan também utilizou estimação de magnitude para verificar graus de ansiedades em 10 mulheres que tinham medo de cobras. O estudo consistiu em apresentar uma cobra viva e uma outra empalhada em distâncias variadas e encontrou uma relação segundo uma função de potência entre as estimativas dos sujeitos e as distâncias das cobras. Embora a cobra viva evocasse substancialmente altas estimativas, os resultados, entre a apresentação da cobra viva e da cobra empalhada, foram similares. Finalmente Sullivan testou a validade do procedimento possibilitando aos sujeitos, a escolha de uma intensidade sonora aversiva (ruído branco) ou de tocar na cobra. Os sujeitos, com medo excessivo de cobra, resistiram intensamente aos níveis altos de ruído branco, porém o teste de significância estatística não apresentou diferenças entre os grupos. Sullivan (1973) elaborou ainda um outro estudo, o qual consistia em emparelhar ruído branco com distância da cobra e para tal, convidou 25 mulheres com medo de cobras. A experiência consistia numa cobra viva engaiolada que se movia até que o sujeito emitisse um intenso ruído branco. Os resultados mostraram novamente, a função de potência como relação entre estimativas de magnitude de ansiedade e distância da cobra e também, uma relação quantitativa da função distância - ansiedade e quantidade de ruído branco emitido.

Diagnósticos psicopatológicos

Numa série de experimentos Stone (1968a) investigou a relação entre julgamentos dos psiquiatras e prognósticos de graus de favorabilidade de psicopatologias, utilizando estimações de magnitude. No primeiro estudo, Stone e Skurdal (1968) descobriram evidências da aplicação da função de potência para prognosticar julgamentos. Para tal, o autor convidou 13 psiquiatras que julgaram os prognósticos utilizando uma escala de categoria de 7 pontos, além da estimação de magnitude. Os resultados indicaram que, este contínuo possui características protéticas, pois a curva obtida apresentou uma concavidade voltada para baixo. Assim, os autores concluíram que as opiniões dos psiquiatras escalonadas desta forma são legítimas e a escala subjetiva obtida tem as propriedades de uma escala de razão. Stone (1969a) verificou a validade das estimações de magnitudes. Para tal, convidou 29 psiquiatras utilizando esta metodologia para julgar prognósticos de 33 pacientes que eram voluntários no estudo de drogas experimentais. Este grupo possuia os diagnósticos das 15 psicoses funcionais descritas no Diagnostic and Statistical Manual, Mental

Disordes (DSM-I, American Psychiatric Association, 1952), além dos diversos graus das enfermidades. Os dados foram relacionados e obteve-se a função de potência como o melhor ajuste. Stone (1968b) estudou também a relação entre os julgamentos dos psiquiatras quanto o grau de severidade das psicopatologias dos pacientes emparelhando índices clínicos comuns (dias de internação, QI e graus de psicoses) da patologia dos mesmos. Os resultados mostraram que, novamente a função de potência se ajustava aos dados coletados. Stone (1969b) validou também a função de potência verificando a suscetibilidade dos pacientes frente ao estresse externo. Para tal, 43 psiquiatras utilizaram estimação de magnitude e 13 utilizaram estimação de categoria numa escala de 7 pontos. Os resultados indicaram baixa correlação entre os dois métodos. Isso ocorreu, provavelmente devido à falta de consenso entre os juízes em delimitar a dimensão do estresse. Porém, esses resultados são mais animadores quando somente as reações esquizofrênicas são consideradas. Stone e Lincheid (1971) encontraram um bom ajustamento da função de potência quando quatro reações depressivas foram excluídas. Stone (1970b) submeteu esses resultados ao método do emparelhamento intermodal para obter validação da função de potência entre a gravidade das psicopatologias e o julgamento dos prognósticos. Se os valores das sensações de dois contínuos protéticos são iguais em vários níveis de intensidade, então o expoente da curva log-log deverá ser a razão entre os expoentes. A razão obtida foi próximo ao expoente predito.

Com a chegada do DSM-II (1968), os psiquiatras foram convidados a elaborar uma escala de razão quando descreviam o papel do estresse, graus de prejuízo, e predisposição da personalidade para morbidade, com os seguintes adjetivos que consistiam em *nenhum, leve, moderada* e *severa*. Stone (1970a) elaborou escalas de categoria e de magnitude e observou que, a curva era levemente côncava e representava uma dimensão protética. O experimento apontou também que, a escala de adjetivos possui praticamente intervalos iguais. Isso introduz a possibilidade de que a quantificação de prontuários médicos de pacientes psiquiátricos que possuam estes diagnósticos.

B. Elaboração de Escalas de Razão

A elaboração das escalas de razão apresenta quatro etapas:

1. Seleção dos atributos
2. Aplicação de teste piloto
3. Aplicação do instrumento contendo a escala
4. Emparelhamento intermodal

1. Seleção dos atributos

A primeira etapa é calcada na coleta de dados, com o propósito de identificar diferentes atributos de fenômenos, os quais consistirão de estímulos para a futura composição da escala. Como por exemplo, diferentes intensidades de gravidade para diferentes tipos de crimes. Em determinadas situações, para verificar quantitativamente as possíveis alterações que diferentes situações ou condições causam a uma amostra, devemos investigar os possíveis atributos que compõem o alvo

da pesquisa. Para tal, deve-se entrevistar um determinado número de pessoas que convivem nessas situações ou condições. Por exemplo, se o pesquisador planeja verificar quais as situações de estresse enfrentado pelos professores no seu dia a dia, sugere-se que sejam feitas entrevistas com esses sujeitos perguntando-lhes como são suas rotinas, seus deveres, seus direitos, relacionamentos com alunos, colegas e superiores, procurando cobrir seu universo profissional, identificando atributos referentes ao ambiente de trabalho.

Nos casos de replicações de escalas existentes, basta utilizar os estímulos constantes nestas escalas, ou adaptar alguns ou ainda acrescentar aqueles que se fizerem necessários.

2. Aplicação de teste piloto

Esta fase tem o propósito de testar se os sujeitos compreenderam as instruções. Para eleger-se o estímulo padrão, é escolhido um estímulo de baixa ou média intensidade.

Selecionados os atributos, pede-se a 30 sujeitos que julguem os mesmos numa escala de 0 a 6, sendo o zero considerado nenhuma ação, 6 total ação e os valores intermediários considerados gradações crescentes de ação do atributo julgado.

Os atributos que tiverem médias em torno de 3 serão os elegíveis ao estímulo padrão. Outra qualidade fundamental do estímulo padrão é a sua familiaridade ao sujeito. Um estímulo padrão ideal pode ser definido como coincidência de 100% entre os sujeitos e terem média 3.

Nas escalas existentes, pode-se utilizar os estímulos padrão propostos pelas mesmas.

3. Aplicação do instrumento contendo a escala

Nesta fase o propósito é de comparação dos métodos: estimação de categorias e estimação de magnitudes.

Coletar os dados instruindo os sujeitos na tarefa a ser feita, solicitando-lhes que tenham o primeiro estímulo ou o estímulo padrão como referência e que assinalem valores proporcionais aos demais atributos (estímulos).

Para analisar os dados das estimativas de magnitude, calcular a média geométrica (MG) e o respectivo desvio-padrão geométrico (DP), inserindo as estimativas na planilha (Excel) (DG) = (MG/MA) XDP. Em se tratando das estimativas de categorias, calcular as médias aritméticas (MA) e os respectivos desvios-padrão (DP).

Em seguida, vem a construção do gráfico, colocando-se no eixo X, as médias geométricas das estimativas de magnitudes e no eixo Y, as médias aritméticas das estimativas de categorias. Se uma curva com a concavidade voltada para baixo é obtida, então, podemos concluir que o contínuo tem características protéticas, pois através desta curva verificamos que nas estimativas de categorias, os sujeitos julgaram diferenças, enquanto que nas estimativas de magnitudes, os sujeitos julgaram razões. Caso a curva apresentada for uma linha reta, então o contínuo possui caracte-

rísticas metatéticas. Reforçando o fato de que o contínuo possui características protéticas, construímos o mesmo tipo de gráfico, apenas agora substituindo as médias geométricas das estimativas de magnitudes (eixo X) pelos logaritmos dessas médias. A tendência é inversão da curva.

Pode-se construir um gráfico que visualiza a Lei de Ekman, no qual a curva deverá ser uma linha reta ascendente, colocando-se as estimativas de magnitudes no eixo X e os desvios padrões geométricos no eixo Y.

Quando a escala apresenta um número maior que 20 estímulos, é conveniente que o primeiro experimento seja somente de estimativas de magnitudes numéricas para avaliação dos mesmos e o segundo experimento seja a comparação entre estimativas de magnitudes e de categorias acima descritos. Para selecionar os estímulos deste experimento, deve-se considerar as médias geométricas resultantes do experimento anterior, considerando-se espaçamento entre elas.

4. Emparelhamento intermodal

Nesta fase o propósito é a validação da escala de razão. A tarefa de emparelhamento é feita utilizando-se duas diferentes modalidades de respostas (como por exemplo, sonoridade, brilhância, comprimento de linhas, força dinamométrica, números, etc).

Coletar os dados utilizando duas diferentes modalidades de respostas, como por exemplo, numérica e comprimento de linha. Utilizar os mesmos estímulos (atributos) para os dois instrumentos, diferenciando apenas as especificidades das instruções.

Para a tarefa de calibração podem ser utilizados os métodos psicofísicos de estimação de magnitudes (Calibração 1) e de produção de magnitudes (Calibração 2). No primeiro método (tarefa de Calibração 1) a tarefa dos participantes consistirá em estimar, por exemplo, os seguintes comprimentos de linhas: 15, 23, 30, 42, 56 e 84 cm. Os participantes devem receber instruções para estimar os seis diferentes comprimentos de linhas assinalando a cada comprimento um número que seja proporcional à sua dimensão aparente. Por exemplo, se o participante julgar que um dado comprimento é duas vezes maior do que aquele primeiramente apresentado, ele deverá assinalar a ele um número duas vezes maior. Se o participante julgar que um outro comprimento parece ter a metade do comprimento apresentado primeiro, ele deverá assinalar a ele um número que seja metade daquele assinalado ao primeiro comprimento apresentado. Os seis diferentes comprimentos de linhas devem ser apresentados em duas séries de seis, uma para treino e a outra usada para análise, os quais devem ser indicados pelo experimentador numa trena, um a um, numa ordem totalmente aleatória para cada participante, sendo que cada um deverá estabelecer 12 estimativas, sendo duas para cada comprimento de linha apresentado. Estímulo padrão e módulo devem ser estabelecidos previamente. No exemplo, o estímulo padrão poderá ser 30 cm e o módulo 100.

No segundo método (tarefa de Calibração 2), a tarefa dos participantes deverá consistir em produzir um comprimento de linha às médias geométricas resultantes das estimativas de magnitudes dadas pelos participantes na tarefa anterior (estimação de magnitude dos comprimentos de linhas: 15, 23, 30, 42, 56 e 84 cm). O

participante deve receber instruções para estimar os diferentes números produzindo um comprimento de linha que seja proporcional à dimensão aparente do número designado. Por exemplo, se o participante julgar que um dado número é duas vezes maior do que aquele primeiramente apresentado, ele deve assinalar a ele um comprimento de linha duas vezes maior. Se o participante julgar que um outro número parece ter a metade da dimensão do número apresentado primeiro, ele deverá assinalar a ele um comprimento de linha que seja metade daquele assinalado ao primeiro número apresentado.

Os seis diferentes números devem ser apresentados em duas séries de seis (uma para treino e a outra usada na análise), os quais serão indicados pelo experimentador, um a um, numa ordem totalmente aleatória para cada participante. Cada participante estabelecerá 12 estimativas, sendo duas para cada número. Estímulo padrão e módulo devem ser estabelecidos previamente. No exemplo, o estímulo padrão poderá ser 100 e o módulo 30 cm.

Para cálculo do expoente a ser predito na tarefa de calibração, utilizar o procedimento do cálculo de equação do gráfico. Em seguida calcular a média dos dois expoentes resultantes dos métodos utilizados na tarefa de calibração.

Para calcular o expoente do contínuo, colocar os logaritmos das médias geométricas das estimativas de comprimento de linhas no eixo X e os logaritmos das médias geométricas das estimativas numéricas no eixo Y. Traçar o gráfico, devendo este ser uma reta, segundo uma função de potência. Para validar psicofisicamente a escala de razão, o expoente obtido deve ser comparado ao predito por Stevens (1975) ou ser comparado ao expoente obtido na tarefa de calibração. Caso o valor médio do expoente obtido esteja dentro da igualdade estatística quando comparado ao expoente predito ou obtido na tarefa de calibração, então podemos considerar válida a escala de razão. No exemplo, os expoentes preditos por Stevens (1975) para número e comprimento de linha é 1,00, logo:

$$\frac{1}{1} \quad \begin{matrix}\text{(expoente predito para comprimentos de linhas)} \\ \text{(expoente predito para números)}\end{matrix} = 1$$

Fazer outro gráfico para verificar se a Lei de Ekman é válida para as duas modalidades de respostas. Nesse gráfico, no eixo X devem ser colocadas as médias geométricas das estimativas da modalidade a ser verificada (no exemplo comprimento de linhas e estimativas numéricas) e no eixo Y as estimativas de variabilidade (erro padrão ou desvio padrão).

Podem ser calculados individualmente os valores de n (expoente), k (constante) e r^2 (coeficiente de determinação).

C. Vantagens da Metodologia Psicofísica Aplicada

As seguintes características da metodologia psicofísica demonstram sua superioridade quando comparada com outros métodos comumente utilizados para escalonar variáveis psicossociais, tais como as escalas de Guttman, Likert, Thurstone e Diferencial Semântico: (1) os sujeitos selecionam livremente as medidas de respostas, (2) o número de participantes pode ser pequeno, (3) as escalas de mensuração

em nível de razão são geradas, e, como conseqüência, aumenta a sensibilidade e o rigor da mensuração pois nesta todas as operações estatísticas e aritméticas são admissíveis, (4) os julgamentos e as escalas produzidas são consistentes e estáveis com coeficientes de fidedignidade (produto-momento) teste-reteste variando de 0,90 a 1,00, (5) os procedimentos de estimação de magnitudes e emparelhamento intermodal são fáceis de serem entendidos e usados pelos pacientes, auxiliares de enfermagem, enfermeiras e médicos, (6) os procedimentos são de baixo custo e não há perda de dados e os mesmos podem ser coletados individual ou coletivamente e, (7) os procedimentos são encarados pelos sujeitos como "jogo", por isso reduzem a fadiga e a monotonia comumente encontradas em outras estratégias. Em resumo, estas vantagens da metodologia psicofísica podem ser enriquecidas pela qualidade dos dados obtidos e a atitude positiva dos participantes.

Para ilustrar estas vantagens pode-se mencionar o estudo de Schepp (1991) que comparou o método de estimação de magnitudes com a escala de Likert nos esforços de mães em manipularem suas energias físicas e emocionais para lidarem com suas crianças hospitalizadas. Os resultados mostraram que as respostas de razão derivadas do método de estimação de magnitude foram substancialmente diferentes daquelas obtidas com a Escala de Likert quanto às intensidades dos estímulos, portanto, sugerindo que estimação de magnitudes é superior em detectar variações com níveis de estímulos elevados, uma vez que com este método a amplitude de respostas é ilimitada. Sennott-Miller, Murdaugh e Hinshaw (1988) registraram que os resultados obtidos através de estimação de magnitude produzem freqüentemente altos coeficientes de fidedignidade (produto-momento) teste-reteste, isto é, ao redor de 0,91.

De outro lado, Wills e Moore (1994) têm argüido que nem sempre o método de estimação de magnitudes é superior ao método de estimação de categorias para escalonar estados subjetivos nas pesquisas. De fato, eles argumentam que em muitas pesquisas o foco de interesse repousa nas diferenças nos estados subjetivos entre grupos de pessoas ou entre indivíduos e, portanto, torna-se importante fazer inferências sobre as diferenças nos estados subjetivos baseados nas diferenças das ordenações destes estados escalonados. Apesar disso, entendemos que é importante os pesquisadores estarem cientes das variabilidades inter e intra sujeitos, oriundas do emprego de qualquer um desses métodos baseados em julgamentos de razão (estimação de magnitudes) ou em julgamentos de diferenças (estimação de categorias).

Referências Bibliográficas

American Psychiatric Association Committee on Nomenclature and Statistics (1952). *Diagnostic and statistical manual: Mental disorders.* (1st ed.) Washington D.C.: American Psychiatric Ass.

Baird, J.C. & Noma, E. (1978). *Fundamentals of scaling and psychophysics*. New York: Wiley.

Birnbaum, M.H. (1992). Predicting healthy from stress: Response to Crandall. *Psychological Science*, *5*, 319-320.

Birnbaum, M.H. & Sotoodeh, Y (1991). Measurement of stress: Scaling the magnitudes of life changes. *Psychological Science*, *4*, 236-242.

Borg, G. (1982). Psychophysical bases of perceived exertion. *Medicine and Science in Sports and Exercise*, *14*, 377-381.

Borg, G. & Ottonson, D. (1986). *The perception of exertion in physical work.* Macmillan: London.

Brennan, E.M., Ryan, E.B., & Dawson, W.E. (1975). Scaling of apparent accentedness by magnitude estimation and sensory modality matching. *Journal of Psychological Research*, *4*, 27-36.

Burdon, J.G.W., Juniper, E.F., Killian, K.J., Hargreave, F.E., & Campbell, E.J.M. (1982). The perception of breathlessness in asthma. *American Review of Respiratory Disease*, *126*, 825-828.

Crandall, C.S. (1992). Psychophysical scaling of stressful life events. *Psychological Science*, *3*, 256-258.

Cross, D.V. (1974). Some technical notes on psychophysical scaling. In H. Moskowitz, B. Scharf, & J.C. Stevens (Eds.), *Sensation and measurement in honor of S.S. Stevens* (pp. 23-36). Dordrecht, the Netherlands: Reidel.

Da Silva, J.A. & Fukusima, S.S. (1989). Constancy of individual exponents for inspired lung volume: A comment on Harver (1987). *Perceptual and Motor Skills*, *68*, 193-194.

Da Silva, J.A. & Macedo, L. (1982). A função-potência na percepção: significado e procedimento de cálculos do expoente. *Arquivos Brasileiros de Psicologia*, *34*, 27-45.

Dawson, W.E. & Brinker, R.P. (1971). Validation of ratio scales of opinion by multimodality matching. *Perception & Psychophysics*, *9*, 413-417.

Dawson, W.E. & Mirando, M.A. (1973). Sensory-modality scale for pronounceability of trigrams and its relation to free-recall learning. *Perceptual and Motor Skills*, *36*, 1219-1224.

Ekman, G. (1962). Measurement of moral judgment: A comparison of scaling methods. *Perceptual and Motor Skills*, *15*, 3-9.

Ekman, G. & Künnapas, T.M. (1960). Note on direct and indirect scaling methods. *Psychological Reports*, *6*, 174.

Ekman, G. & Künnapas, T.M. (1962a). Scales of aesthetic value. *Perceptual and Motor Skills*, *14*, 19-26.

Ekman, G. & Künnapas, T.M. (1962b). Measurement of aesthetic value by "direct" and "indirect" methods. *Scandinavian Journal Psychology*, *3*, 3-12.

Ekman, G. & Künnapas, T.M. (1963). A further of direct and indirect scaling methods. *Scandinavian Journal of Psychology*, *4*, 77-80.

Ellis, L.W. & Fucci, D.J. (1991). Magnitude estimation scaling of speech intelligibility: Effects of listeners' experience and semantic-syntactic context. *Perceptual and Motor Skills*, *73*, 295-305.

Engelman, A. (1966). A lei de potência de Stevens: Um caso de constância perceptiva? *Jornal Brasileiro de Psicologia*, *3*, 19-48.

Faleiros Sousa, F.A.E. & da Silva, J.A. (1996). Uso e aplicação da metodologia psicofísica na pesquisa em enfermagem. *Revista Latino-Americana de Enfermagem*, *4*,147-178.

Faleiros Sousa, F.A.E. *Métrica do consenso social e clínico: um enfoque experimental*, Ribeirão Preto, 1997, 263 p. Tese (Livre docência), Escola de Enfermagem de Ribeirão Preto da Universidade de São Paulo.

Faleiros Sousa, F.A.E. *Prestígio profissional do enfermeiro: um enfoque da Psicofísica Social*, Ribeirão Preto, 1993, 197p., Tese (doutorado), Escola de Enfermagem de Ribeirão Preto da Universidade de São Paulo.

Fucci, D.J., Ellis, L.W., & Petrosino, L. (1990). Speech clarity/intelligibility. *Journal of Experimental Psychology*, *58*, 405-413.

Fukusima, S.S., Ribeiro, G., & Da Silva, J.A. (1988). Cálculo da função de potência de Stevens para microcomputadores. *Psicologia: Teoria e Pesquisa*, *4*, 96-101.

Geller, D. & Margolis, R.H. (1984). Magnitude estimation of loudness I: Application to hearing aid selection. *Journal of Speech and Hearing Research*, *27*, 20-27.

Gescheider, G.A. (1988). Psychophysical scaling. *Annual Review of Psychology*, *39*, 169-200.

Gescheider, G.A. (1997). *Psychophysics: The fundamentals*. Mahwah, New Jersey: LEA.

Gottfried, S.B., Redline, S., & Altose, M.D. (1985). Respiratory sensation in chronic obstructive pulmonary disease. *American Review of Respiratory Disease*, *132*, 954-959.

Hamblin, R.H. (1971a). Mathematical experimentation and sociological theory: A critical analysis. *Sociometry*, *34*, 423-452.

Hamblin, R.L. (1971b). Ratio measurement for the social sciences. *Social Forces*, *50*, 191-206.

Hamblin, R.L. & Smith, C.R. (1966). Values, status and professors. *Sociometry*, *29*,183-196.

Hardin, C. & Birnbaum, M.C. (1990). Malleability of "ratio" judgments of occupational prestige. *American Journal of Psychology*, *103*, 1-20.

Harver, A. (1987). Constancy of individual exponents for category production of inspired lung volume. *Perceptual and Motor Skills*, *65*, 779-785.

Harver, A. & Kotses, H. (1987). Perception of static respiratory forces in young and old subjects. *Perception & Psychophysics*, *41*, 449-454.

Harver, A., Tenney, S.M., & Baird, J.C. (1986). A cautionary note on the interpretation of the power law for respiratory effort. *American Review of Respiratory Disease*, *133*, 341-342.

Holmes, T.H. & Rahe, R.H. (1967). The social readjustment rating scale. *Journal of Psychosomatic Research*, *11*, 213-218.

Indow, T. (1961). An example of motivation research applied to product design. *Chosa, To Gijutsu*, *102*, 45-60.

Kemp, S. (1988). Magnitude estimation of the utility of nonmonetary items. *Bulletin of the Psychonomic Society*, *26*, 544-547.

Kemp, S. (1991). Magnitude estimation of the utility of public goods. *Journal of Applied Psychology*, *76*, 533-540.

Killian, K.J., Mahutte, C.K., & Campbell, E.J.M. (1981). Magnitude scaling of externally added loads to breathing. *American Review of Respiratory Disease*, *123*, 12-15.

Knight, K.K. & Margolis, R.H. (1984). Magnitude estimation of loudness II: Loudness perception in presbycusic listeners. *Journal of Speech and Hearing Research*, *27*, 28-32.

Kohn, S.D. (1965). Scaling musical preferences. *Journal of Experimental Psychology*, *70*, 79-82.

Krueger, L.E. (1989). Reconciling Fechner and Stevens: Toward a unified psychophysical law. *Behavioral and Brain Science*, *12*, 251-320.

Künnapas, T.M. & Wikstroem, I. (1963). Measurement of occupational preferences: A comparison of scaling methods. *Perceptual and Motor Skills*, *17*, 611-624.

Le Blanc, P., Bowie, D. M., Summers, E., Jones, N. L., & Killian, K. J. (1986). Breathlessness and exercise in patients with cardiorespiratory disease. *American Review of Respiratory Disease*, *133*, 21-25.

Lodge, M. (1982). *Magnitude scaling: Quantitative measurement of opinions*. Beverly Hills, CA: Sage Publications.

Lodge, M., Cross, D., Tursky, B., & Tanenhaus, J. (1975). The psychophysical scaling and validation of a political support scale. *American Journal of Political Science*, *19*, 611-649.

Lodge, M., Cross, D., Tursky, B., Tanenhaus, J., & Reeder, R. (1976). The psychophysical scaling of political support in the "real world". *Political Methodology*, *2*, 159-182.

Lodge, M., Cross, D., Tursky, B., Foley, M.A., & Foley, H. (1976). Calibration and cross-modal validation of ratio scales of political opinion in survey research. *Social Science Research*, *5*, 325-347.

Mahler, D.A., Rosiello, R.A., Harver, A. Lentine, T., McGovern, J. F., & Daubenspeck, J. A. (1987). Comparison of clinical dyspnea ratings and psychophysical measurements of respiratory sensation in obstructive airway disease. *American Review of Respiratory Disease*, *135*, 1229-1233.

Masuda, M. & Holmes, T. H. (1967). Magnitude estimations of social readjustments. *Journal of Psychosomatic Research*, *11*, 219-225.

Muheenkamp, A.F., Gress, L.D., & Flood, M.A. (1975). Perception of life change events by the elderly. *Nursing Research*, *24*, 109-113.

Nield, M., Kim, M.J., & Patel, M. (1989). Use of magnitude estimation for estimating the parameters of dyspnea. *Nursing Research*, *38*, 77-80.

Nield, M. & Kim, M.J. (1991). The reliability of magnitude estimation for dyspnea measurement. *Nursing Research*, *40*, 17-19.

Perloe, S.I. (1963). The relation between category-rating and magnitude-estimation judgments of occupational prestige. *American Journal of Psychology, 76,* 395-403.

Poulton, E.C. (1968). The new psychophysics: Six models for magnitude estimation. *Psychological Bulletin, 69,* 203-214.

Rahe, R.H., Meyer, M., Smith, M., Kjaer, G., & Holmes, T.H. (1964). Social stress and illness onset. *Journal of Psychosomatic Research, 8,* 35-44.

Ruch, L.O. & Holmes, T.H. (1971). Scaling of life change: Comparison of direct and indirect methods. *Journal of Psychosomatic Research, 15,* 221-227.

Schepp, K.G. (1991). Factors influencing the coping effort of mothers of hospitalized children. *Nursing Research, 40,* 42-46.

Schiavetti, N., Metz, D.E., & Sitler, R.W. (1981). Construct validity of direct magnitude estimation and interval scaling of speech intelligibility: Evidence from a study of the hearing impaired. *Journal of Speech and Hearing Research, 24,* 441-445.

Sellin, J.T. & Wolfgang, M.E. (1964). *The measurement of delinquency.* New York: Wiley.

Shinn, Jr., A.M. (1969). An application of psychophysical scaling techniques to the measurement of national power. *Journal of Politics, 31,* 932-951.

Stevens, S.S. (1959). Cross-modality validation of subjective scales for loudness, vibration, and electric shock. *Journal of Experimental Psychology, 57,* 201-209.

Stevens, S.S. (1960). The psychophysics of sensory function. *American Scientist, 48,* 226-253.

Stevens, S.S. (1966a). A metric for the social consensus. *Science, 151,* 530-541.

Stevens, S.S. (1971). Issues in psychophysical measurement. *Psychological Review, 78,* 426-450.

Stevens, S.S. (1974). Perceptual magnitude and its measurement. In E.C. Carterette & M. Friedman (Eds.), *Handbook of perception* (Vol. 2, pp. 361-389). New York: Academic Press.

Stevens, S.S. (1975). *Psychophysics: Introduction to its perceptual, neural and social prospects.* In G. Stevens (Ed.), New York: Wiley.

Stevens, J.C., Mack, J.D., & Stevens S.S. (1960). Growth of sensation on seven continua as measured by force of handgrip. *Journal of Experimental Psychology, 59,* 60-67.

Stevens, S.S. & Guirao, M. (1962). Loudness, reciprocality, and partition scales, *Journal of the Acoustical Society of America, 34,* 1466-1471.

Stevens, S.S. & Greenbaum, H.B. (1966). Regression effect in psychophysical judgment. *Perception & Psychophysics, 1,* 439-446.

Stone, L.A.(1968a). Clinical psychophysics. *Studia Psychologica, 10,* 161-173.

Stone, L.A.(1968b). Bases for psychiatric impairment severity judgements: Psychophysical power functions? *Studia Psychologica, 10,* 194-199.

Stone, L.A.(1969a). Bases for psychiatric prognostic favorability judgements: Psychophysical power functions? *Behavior Science, 14,* 133-137.

Stone, L.A.(1969b). Psychiatric's judgmental evaluations of susceptibility to external stress for selected disorder classification stimuli. *Journal of Clinical Psychology, 25,* 21-26.

Stone, L.A.(1970a). Magnitude estimation and numerical category scale evaluations of category scale adjectival stimuli on three clinical judgmental continua. *Journal of Clinical Psychology, 26,* 24-27.

Stone, L.A.(1970b). A law of clinical judgment: A psychological mechanism based on the logic of psychophysics. *Journal of Clinical Psychology, 26,* 312-317.

Stone, L.A. & Lincheid, T.R.(1971). Another law of clinical judgment: A psychological mechanism based on the logic of psycho-dynamics. *Psychological Reports, 28,* 851-855.

Stone, L.A. & Skurdal, M.A. (1968). Judged prognosis for functional psychosis disorder classifications: A prothetic continuum. *Journal of Consulting and Clinical Psychology, 32,* 469-472.

Sullivan, R. (1969). Subjective matching of anxiety to intensities of white noise. *Journal of Abnormal Psychology, 74,* 646-650.

Sullivan, R. (1970). Magnitude estimation of anxiety. *Psychonomics Science, 21*(4), 209-211.

Sullivan, R. (1971). Magnitude estimation and relative aversiveness of anxiety: Phobia. *Journal of Abnormal Psychology, 78,* 266-271.

Sullivan, R. (1973). A method for scaling its relative magnitude and aversiveness. *Journal of Abnormal Psychology, 82,* 483-490.

Thalmann, R. (1965). Cross-modality matching in the study of abnormal loudness functions. *Laryngoscope, 75,* 1708-1726.

Tryon, W.W. (1977). Psychophysical Scaling and Hierarchy Construction. *Journal of Behaviorism, Therapy & Experimental Psychiatry, 8,* 53-56.

Volicer, B. J. & Bohannon, M. W. (1975). A hospital stress rating scale. *Nursing Research, 24,* 352-364.

Warren, R.M. & Warren, R.P. (1963). As critique of S.S. Stevens "new psychophysics". *Perceptual and Motor Skills, 16,* 797-810.

Wills, C.E. & Moore, C.F. (1994). A controversy in scaling of subjective states: Magnitude estimation versus category rating methods. *Research in Nursing & Health, 17,* 231-237.

Wolpe, J. (1969). Quantitative relationships in the systematic desensitization of phobias. *American Journal of Psychiatry, 119,* 1062-1068.

Wyler, A.R., Masuda, M., & Holmes, T.H. (1968). Seriousness of illness rating scale. *Journal of Psychosomatic Research, 11,* 363-374.

Wyler, A.R., Masuda, M., & Holmes, T.H. (1970). The seriousness of illness rating scale: Reproducibility. *Journal of Psychosomatic Research, 14,* 59-64.

Wyler, A.R., Masuda, M., & Holmes, T.H. (1971). Magnitude of life events and seriousness of illness. *Psychosomatic Medicine, 33,* 115-122.

ESCALAS PSICOMÉTRICAS

Luiz Pasquali

A. Introdução

1. Conceituação de Escala

A medida escalar constitui uma das várias formas que a medida psicométrica pode assumir. Nesta se incluem os testes psicológicos, os inventários, as escalas, etc. Contudo, as medidas escalares propriamente ditas são mais utilizadas na Psicologia Social, especificamente no estudo das atitudes, e também no campo da Personalidade, com o intuito de medir traços de personalidade (como, por exemplo, o inventário de Comrey: Escalas de Personalidade de Comrey). Elas se distinguem dos testes e inventários, porque aqueles são de uso mais corrente na avaliação das aptidões (onde há respostas certas e erradas) e estes no campo da personalidade e da psicopatologia. Além disso, os testes e os inventários, em confronto com as escalas, se apresentam como medidas para as quais existem normas de interpretação, ao passo que para as escalas comumente não são elaboradas tais normas. Na verdade, diferenças essenciais entre estes vários tipos de medidas psicométricas não existem. Há mesmo dúvidas quanto a existirem diferenças importantes entre escalas psicométricas e escalas psicofísicas. A distinção, neste último caso, talvez ainda faça sentido. A escala psicofísica visa escalonar estímulos físicos (através de medida fundamental) que corresponderia ou produziria uma escala intervalar psicológica (escala de resposta), sendo as duas relacionadas por alguma lei psicofísica. A escala psicométrica visa escalonar estímulos que expressam um construto psicológico e seria mais neste sentido restrito que se usaria mais comumente o conceito de escala psicométrica. Mas estas distinções se tornam muito tênues, porque, afinal, sempre se escalonam estímulos (itens) observáveis.

Também, a expressão escala é utilizada de múltiplas formas: para designar o nível métrico da medida (escala ordinal, intervalar, etc.); para designar um contínuo de números (escala numérica de 5 pontos, por exemplo); para designar os próprios itens de um instrumento, como no caso do diferencial semântico, onde cada item é chamado de escala; para designar diferentes técnicas de construção e uso de instrumentos psicológicos de medida de atitudes (como, escala tipo Thurstone, tipo Likert, etc.). Todos estes são usos legítimos da palavra escala e, mesmo, não há con-

tradições em tais usos. Embora eles possam trazer algumas dificuldades, normalmente não produzem ambigüidades no tipo de escala que se está falando. O termo, na verdade, originalmente se refere ao fato de que ao se proceder a uma medida de um atributo empírico, surge uma série de números ordenados à qual é dado o nome de escala numérica. Assim, qualquer medida resultaria numa escala. No caso presente, entretanto, escala é utilizada como uma forma ou técnica de se fazer a medida, especialmente na área das atitudes, como se verá a seguir.

De qualquer forma, escala se refere a instrumento de medida em Psicologia. Ele se caracteriza por ser composto por uma seqüência de números, tipicamente obedecendo os axiomas de ordem dos números, isto é, a seqüência é, pelo menos, do tipo monotônica crescente (ou decrescente, se quiser). O que é que estes números representam? Como se trata de medida, consequentemente eles se referem a algum aspecto da realidade, seja física ou mental ou outra, e desejam indicar diferentes magnitudes de uma propriedade ou atributo desta realidade.

Estas observações permitem que você olhe para esta escala de números ou com olhos de estatístico ou com olhos de cientista, no nosso caso, olhos de psicólogo. O estatístico se pergunta que axiomas do número esta escala monotônica crescente salvaguarda e o psicólogo/cientista se pergunta que atributo da realidade ou psicológico ela está descrevendo. Ambas as perguntas são importantes, pois da primeira surge como esta escala deve ser manipulada nas análises estatísticas e, da segunda, surgem as inferências de ordem teórica sobre o traço ou atributo que a escala está medindo. A primeira pergunta é respondida dentro do tema das *escalas de números* (ordinal, intervalar, etc.) e a segunda é respondida pelos vários tipos de escalas de atitude, por exemplo, que é o objeto deste capítulo. Antes, porém, vamos rever brevemente a história das escalas de números, já que elas tem incidência direta no tratamento dos dados que surgem das escalas de atitude.

2. As Escalas de Números

Dependendo da quantidade de axiomas do número que a medida salva, resultam vários níveis de medida, chamadas as escalas de medida ou escalas de números. São três os axiomas básicos do número: identidade, ordem e aditividade. O último apresenta dois aspectos úteis para o presente problema: origem e intervalo ou distância. Quanto mais axiomas do número a medida salvaguardar, maior será o seu nível, isto é, mais ela se aproxima da escala numérica ou métrica e maior será o isomorfismo entre o número e as operações empíricas. Assim, podemos considerar cinco elementos numéricos para definir o nível da medida: identidade, ordem, intervalo, origem, e unidade de medida. Destes cinco elementos, os mais discriminativos dos níveis são a origem e o intervalo, dado que a ordem é uma condição necessária para que realmente haja medida. Se a medida somente salva a identidade do número, na verdade não se trata de medida, mas sim de classificação e contagem. Neste caso (escala nominal), os números não são atribuídos a atributos dos objetos, mas o próprio objeto é identificado por rótulo numérico, como, por exemplo, dando 1 aos sujeitos de sexo masculino e 2 aos do feminino. Este rótulo nem precisaria ser numérico dado que não importa que símbolo ou rabisco pode ser utilizado com a mesma função de distinguir objetos um do outro ou classe de objetos de outra classe. A

única condição necessária é que se salvaguarde a identidade do símbolo, isto é, um mesmo símbolo não pode ser duplicado para identificar objetos diferentes, como também diferentes símbolos não podem ser usados para identificar objetos idênticos. Embora não estejamos neste caso medindo, a escala numérica que resulta desta rotulação adquire direito ao nome escala, dado que ela corresponde em parte à definição de medida que reza "medir é atribuir números às coisas empíricas".

O esquema abaixo ilustra como se originam as várias escalas de medida:

		Origem	
		Não-natural	Natural
Intervalo	Não-igual	Ordinal	Ordinal
	Igual	Intervalar	Razão

A Tabela 5-1 sumaria as características de cada escala.

Tabela 5-1. Características das escalas numéricas de medida

Escala	Axiomas Salvos	Invariâncias	Liberdades	Transformações Permitidas	Estatísticas Apropriadas
Nominal	- identidade		- ordem - intervalo - origem - unidade	Permutação (troca 1 por 1)	Freqüências: f, %, p, Mo, qui2, C
Ordinal	- identidade - ordem	- ordem	- intervalo - origem - unidade	Monotônica crescente (isotonia)	Não-paramétricas: Md, rs, U, etc.
Intervalar	- identidade - ordem - aditividade	- ordem - intervalo	- origem - unidade	Linear de tipo y=a+bx	Paramétricas: M, DP, r, t, f, etc.
Razão	- identidade - ordem - aditividade	- ordem - intervalo - origem	- unidade	Linear de tipo y = bx (similaridade)	M geométrica, Coef.variação, Logaritmos

Uma escala numérica pode ser transformada numa outra equivalente se forem respeitados os elementos da invariância nesta transformação. Uma escala de maior nível pode utilizar as operações estatísticas de uma escala inferior, mas perde informação dado que as estatísticas próprias de uma escala inferior são menos eficientes, isto é, são menos robustas. Não é permitido (é erro) utilizar estatísticas de uma escala de nível superior numa inferior, dado que esta não satisfaz os requisitos necessários para se utilizar de procedimentos estatísticos superiores. São chamados de paramétricos os procedimentos estatísticos da escala intervalar porque os números nela possuem caráter métrico, isto é, são adicionáveis, enquanto os não-paramétricos não são métricos, dado que representam somente postos e não quantidades somáveis.

3. Escala Psicológica

Aqui existe uma relativa confusão entre escalas psicométricas e psicofísicas, que procuraremos explicar mais adiante. Por ora, vamos nos concentrar no que elas representam de um ponto de vista da Psicologia. Um esquema de como concretamente estas escalas surgem talvez possa elucidar esta questão. Há várias maneiras de apresentar um tal esquema; contudo, sempre se trata da situação em que há respostas de um organismo diante de um estímulo, ou mais tipicamente, de vários organismos (ou sujeitos) diante de vários estímulos. Por enquanto vamos considerar a situação das escalas unidimensionais e, por isso, os estímulos de que estamos falando são considerados como se referindo a um e mesmo atributo de uma realidade (ou de um sistema real).

Um tal esquema pode ser ilustrado como segue:

Sujeito (i)	Estímulo (j)				Soma
	a	b	c	...	
1	R_{1a}				
2					
3			R_{ij}		
...					
Soma					

Neste esquema, o R_{ij} representa a resposta dada pelo sujeito *i* ao estímulo *j*; por exemplo, R_{1a} é a resposta dada pelo sujeito 1 ao estímulo *a*. Agora, você pode estar interessado em produzir uma escala de estímulos, por exemplo. Neste caso, você somaria sobre as linhas, obtendo um escore total para cada estímulo *j*, com base no qual os diferentes estímulos poderiam ser escalonados numa escala crescente definida pela grandeza do seu respectivo escore. Os vários sujeitos, nesta situação, servirão apenas de replicações, isto é, bastava um sujeito para se conseguir tal escala; contudo se trabalha com muitos sujeitos para que o escore seja mais estável e tipicamente se toma, como escore do estímulo, a média ou a mediana das respostas de todos os sujeitos para cada item. Esta é a preocupação ou a tecnologia das escalas de Thurstone.

Ou você está interessado em escalonar os sujeitos; neste caso, você obtém um escore para cada sujeito, somando sobre as colunas, isto é, sobre os itens, tornando-se, assim, estes apenas replicações. Esta é a tecnologia das escalas de Likert.

Ou ainda, você pode estar interessado em escalonar ambos, os itens e os sujeitos, fazendo a suposição que a resposta do sujeito depende tanto das características dos itens quanto dos próprios sujeitos. Esta é a tecnologia dos escalogramas de Guttman.

Assim, vemos que destas três abordagens surgem tipos de escalas psicológicas diferentes, dado que fazem suposições diferentes. De fato, Guttman considera a resposta do sujeito como resultante tanto dos estímulos quanto dos sujeitos, enquanto Thurstone considera que a resposta do sujeito depende exclusivamente do estímulo, ao passo que Likert faz o contrário, a resposta depende exclusivamente do sujeito. Temos, então, as seguintes equações:

(5.1) $R = f(S,E)$ Guttman
$R = f(E)$ Thurstone
$R = f(S)$ Likert

onde,

S = sujeito
R = resposta
E = estímulo.

Evidentemente, todas estas posições receberam e recebem críticas, mas sobretudo a de Thurstone tem sido avaliada como pouco defensável (Luce, 1977); a de Guttman, que parece a mais razoável, tem encontrado enormes dificuldades práticas no escalonamento dos estímulos. Por isso, as escalas de Thurstone e mesmo de Guttman são raramente utilizadas hoje em dia em Psicologia, deixando o lugar para as escalas de Likert, que são quase abusivamente utilizadas, sobretudo em Psicologia Social.

4. Escala Psicométrica e Escala Psicofísica

Dentro deste esquema, em que se diferenciaria agora uma escala psicométrica de uma escala psicofísica? Escalonar estímulos seria uma escala psicofísica, porque escalona realidades físicas? E escalonar sujeitos seria escala psicométrica porque escalona respostas de organismos? Mas a resposta de organismos também é uma realidade física ou fisiológica, se quiser. Neste caso, a diferença seria irrelevante. Pode-se, contudo, manter a diferenciação entre estes tipos de escalas, dependendo do enfoque epistemológico que implicitamente está por trás de concepções desta natureza. Dentro de um enfoque epistemológico behaviorista (materialista), onde o objeto da Psicologia é exclusivamente o comportamento (físico), a distinção entre escala psicométrica e psicofísica não aparece. Se, contudo, considerarmos os estímulos (físicos) como representantes de traços latentes, então seria possível se diferenciar escala psicométrica, que trabalha com estímulos representantes de traços psicológicos, de uma escala psicofísica que trabalha com estímulos físicos simplesmente. Mas esta posição supõe uma visão dualista, digamos cognitivista, da Psicologia, que estudaria os traços latentes via representação física no comportamento do organismo. Neste caso, as próprias escalas de Thurstone seriam escalas psicométricas, como aliás ele próprio as chama, dado que os estímulos são estabelecidos via "julgamento".

De fato, a fórmula de Thurstone, R = f (E) pode assumir significado duplo, a saber, uma psicofísica e outra psicométrica, dependendo de como é definido o elemento *E* (estímulo). Na expressão psicofísica, o *E* é entendido como dimensão física: os estímulos (E) se situam num contínuo físico. Na expressão psicométrica, o *E* é entendido como dimensão não-física, subjetiva, psicológica, a saber, o julgamento (J); neste caso os *E* se situam num contínuo psicológico e não mais físico. Uma escala composta de *E* entendidos nesta última maneira Thurstone chama de escala psicométrica em oposição a escala psicofísica onde os *E* se situam no contínuo físico.

Desta maneira de conceber os *E*, também resultam técnicas diferentes de medida dos mesmos. Em psicofísica, os *E* são medidos via mensuração fundamental ou derivada, ao passo que na medida psicométrica, eles são mensurados indi-

retamente via medida por teoria (vide Pasquali, 1996, 1998), isto é, através das respostas subjetivas dos sujeitos que expressam comportamentalmente os julgamentos por eles feitos com referência aos estímulos. A estes julgamentos Thurstone deu o nome de "processos discriminantes", como a seguir será explicado, que na psicofísica moderna são chamados de "representação do sinal" ou de "variável aleatória de decisão" (Luce, 1977).

As escalas psicofísicas visam verificar e descrever a correlação que existe entre estímulos físicos (som, peso, tamanho, etc.) e a resposta do sujeito. Mais especificamente, qual é o mínimo valor do estímulo que é capaz de produzir uma resposta no organismo (limiar absoluto) e qual é o mínimo de acréscimo no estímulo necessário para produzir no organismo uma resposta diferente da anterior (limiar diferencial). A determinação do limiar absoluto se faz em termos de 50% de percepção de um dado estímulo: o nível de estímulo que é percebido em 50% das vezes é considerado o limiar absoluto ou nível 0 (zero - inicial) da escala de resposta. Para a determinação dos limiares diferenciais, várias leis foram apresentadas na história da Psicologia. Weber (Stevens, 1951) concebeu a lei da constante: para produzir uma resposta diferente da anterior, o estímulo deve ser aumentado por uma constante (k) que deve ser determinada empiricamente para cada modalidade de estímulo (peso, som, etc.). Como logo se percebeu que esta lei não correspondia muito à observação dos fatos, Fechner (Stevens, 1951) apresentou uma lei logarítmica, na qual a resposta depende de uma constante, diferente para cada modalidade de estímulo, a qual multiplica o logaritmo do estímulo; isto é, para produzir uma resposta diferente da anterior, o estímulo tem que aumentar logaritmicamente: a resposta aumenta aritmeticamente e o estímulo geometricamente. Outras leis vieram substituir a de Fechner (Stevens, 1951; Guilford, 1954). Stevens de fato demonstrou que alguns pressupostos de Fechner não podiam ser mantidos e introduziu novos procedimentos que vieram a se caracterizar como a lei da potência .

Uma exposição detalhada das medidas psicofísicas vai além da intenção deste capítulo. Para tal informação devem ser consultados os trabalhos de Stevens (1951) e Guilford (1954) e o capítulo 4 do presente livro.

Dar-se-ão mais detalhes na exposição das escalas propriamente psicométricas a seguir. Na apresentação dos vários tipos de escalas, três níveis de preocupação devem ser levados em conta: os procedimentos teóricos, os procedimentos empíricos (experimentais) e os procedimentos analíticos, os quais discriminariam diferentes tipos de escalas psicológicas, procedimentos que foram expostos e justificados na capítulo 2 deste livro. Um manual prático para trabalhar com escalas psicométricas é o livro de A.L. Edwards (1957), *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts, Inc.

4. Escala Unidimensional e Escala Multidimensional

Uma escala é dita unidimensional se ela expressa, através dos números, apenas uma dimensão subjacente. Assim, se o psicólogo deseja medir apenas um traço latente, como ansiedade, por exemplo, ele iria construir uma escala unidimensional. Ou se ele deseja medir vários traços latentes, mas um independentemente do outro, ele iria construir tantas escalas quantos são os traços latentes que

deseja medir. Por exemplo, o inventário de personalidade IFP (Pasquali, 1997) mede 16 traços de personalidade, mas continua sendo um aglomerado de escalas unidimensionais, porque existe um escala independente para medir cada um dos traços isoladamente; o teste apenas agrupa num único inventário todas estas escalas unidimensionais. Um tal inventário pode ser chamado de escalas multifatoriais, mas não de inventário ou escala multidimensional. A distinção se tornará clara logo em seguida.

Por outro lado, se se diz que a escala é multidimensional, tal asserção implica que há mais de um traço latente afetando as respostas dos sujeitos num dado conjunto de itens, isto é, uma série de observações, expressas por números (escala), têm mais de um traço latente como causa. Uma escala unidimensional (seja de estímulos ou de pessoas) posiciona estes estímulos ou pessoas num espaço unidimensional, ao passo que uma escala multidimensional coloca um objeto num espaço multidimensional. Assim, o IFP, por exemplo, coloca 16 traços latentes em 16 espaços unidimensionais, pois os 16 vetores não estão relacionados entre si. Diferentemente, a escala multidimensional se pergunta quantos traços latentes são necessários para expressar simultaneamente um dado objeto. Intuitivamente uma escala multidimensional é muito simpática, porque parece claro que qualquer resposta do sujeito é determinada por mais de uma causa. Por exemplo, se você quer saber qual de duas pessoas é mais simpática, com toda a certeza os sujeitos ao julgá-las irão compará-las sob vários pontos de vistas, como beleza, juventude, simpatia, atração etc. Então a resposta que eles dão não é determinada por um único traço. Como esta situação é típica em Psicologia, seguiria que as escalas multidimensionais seriam o modelo de escalas a ser utilizado pelos psicólogos. Entretanto, duas observações são importantes: 1) quando se diz que a escala é unidimensional, basta supor que ela esteja medindo *predominantemente* um único traço latente, sendo os demais considerados secundários, onde a variância que estes produzem nos itens fica inserida na variância específica de cada item, restando como variância comum aquela que os itens têm em conjunto e estando referida ao traço latente que a escala quer especificamente medir; 2) em segundo lugar, as escalas multidimensionais, além de extremamente complexas, utilizam conceitos que são tipicamente ambíguos e, consequentemente, dificilmente tratáveis em pesquisas que visam desenvolver a teoria psicológica (Clausen & Van Horn, 1977; McIver & Carmines, 1981). Por estas razões e por serem de simples intelecção, as escalas unidimensionais são ainda as mais utilizadas na pesquisa psicológica, especialmente na pesquisa social.

B. O Enfoque de Thurstone

B.1 Procedimentos Teóricos

Caracterizando o *pólo teórico* de sua posição, Thurstone (1927) introduziu o conceito de contínuo psicológico em oposição ao contínuo físico da psicofísica. A diferença é a seguinte: suponha 10 objetos de igual tamanho mas com pesos diferentes. Estes objetos podem ser ordenados pelo peso de duas maneiras. Primeiro, pode-se usar uma balança e ordenar os objetos pelo seu peso real, produzindo um contínuo físico (através de medida fundamental); mas, segundo, pode-se

também pedir a indivíduos, na falta de uma balança, para avaliar subjetivamente o peso destes objetos e ordená-los do mais leve ao mais pesado e esta ordenação constitui um contínuo psicológico de pesos. Esta ordenação psicológica pode ser feita pelos sujeitos comparando os 10 objetos dois a dois até se chegar à ordem final. Assim, a ordenação dos pesos é feita, não fisicamente, mas via julgamento. Agora, esta é tipicamente a situação quando queremos medir estímulos psicológicos ou traços latentes, para os quais não temos metros ou balanças físicas. Com base neste raciocínio, Thurstone (1927) desenvolveu a lei do julgamento comparativo, que pode ser considerada como introduzindo o conceito de métodos de escalagem psicológica (ou métodos psicométricos em sentido estrito) em oposição aos métodos psicofísicos.

Neste caso, os estímulos (E) não são mais expressos num contínuo físico, mas num contínuo não-físico, subjetivo, chamado precisamente de contínuo psicológico. Os *E* são, então, reações do sujeito a estímulos físicos. Thurstone fez ainda a suposição de que o mesmo estímulo físico irá produzir a mesma reação média ("processo discriminante modal", que será explicado a seguir) no mesmo sujeito, suposição não aceitável na psicofísica moderna (Luce, 1977), porque com isso o estímulo se tornaria uma constante na equação. Aliás, de acordo com esta suposição, a equação de Thurstone, R = f (E), deveria de fato ser expressa como R = E f(J), onde *J* é o julgamento que produz os estímulos do contínuo psicológico, como se ilustra na figura 5-1.

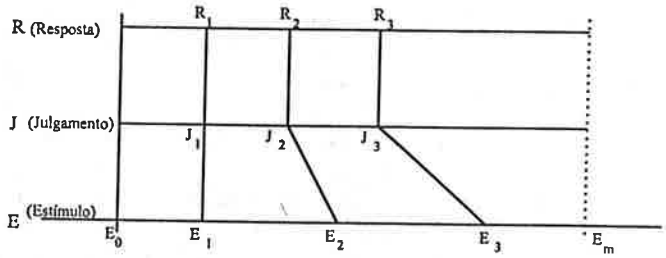


Fig. 5-1. Os contínuos da teoria de Thurstone

A figura 5-1 mostra que

- há três contínuos: o físico (E), o da resposta (R) e o do julgamento (J)
- o R e o J são isomorfos, isto é, a magnitude do J é igual à da R
- As reações subjetivas (J e R) somente cobrem certa extensão das magnitudes de E, isto é, entre E_0 e E_m.

Medindo-se os *E* e as *R* diretamente e estabelecendo a função que os relaciona, temos uma escala psicofísica. Medindo-se as *R* diretamente e os *E* indiretamente, isto é, via *J*, temos uma escala psicométrica, visto que as magnitudes dos *E* não são dadas via metro, balança, etc., mas via reação subjetiva dos sujeitos (os J). Neste caso, as *R* dependem dos *J* e não diretamente dos *E*. Thurstone estabeleceu uma lei que relaciona as *R* e os *E* avaliados via *J*, chamando a esta lei, a lei dos julgamentos comparativos.

A lei do julgamento comparativo se explicita assim: ao comparar dois estímulos 'i' e 'j' para decidir qual deles é maior (ou 'mais do que' em algum atributo dado, como peso, por exemplo), o sujeito tem que fazer três julgamentos. Primeiro ele tem que avaliar o estímulo 'i', depois o estímulo 'j' e, finalmente, a diferença j > i. Ao avaliar os dois estímulos individuais, o sujeito reage a cada um com uma representação subjetiva deles, produzindo um processo discriminante ("discriminal process") para cada um e ao avaliar a diferença entre os dois estímulos, ele produz uma diferença discriminante ("discriminal difference"). Contudo, ao fazer esses julgamentos em ocasiões diferentes, o mesmo sujeito não produz os mesmos processos discriminantes, de sorte que daí resulta uma variabilidade chamada dispersão discriminante em torno de um processo discriminante modal que corresponde à média dos vários processos discriminantes com referência a cada estímulo individual ($\overline{R}_i$ e $\overline{R}_j$). Assim, para cada estímulo, sobre o qual existe uma série de julgamentos (muitos sujeitos avaliando o mesmo estímulo ou o mesmo sujeito avaliando o estímulo em muitas ocasiões diferentes: *procedimentos experimentais*), temos um processo discriminante modal e uma dispersão discriminante, isto é, a média e o desvio padrão, dado que os processos discriminantes se distribuem normalmente (suposição razoável), como aparece na figura 5-2. Recorde, ainda, que numa distribuição normal a média, a mediana e a moda são idênticas. O mesmo acontece com a distribuição das diferenças (j - i) que irão produzir uma distribuição normal com uma média ou diferença discriminante modal ($\overline{R}_{j-i}$) com seu respectivo desvio padrão ou dispersão discriminante (σ_{j-i}).

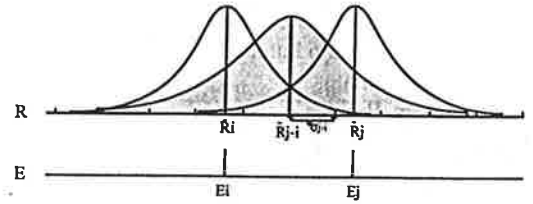


Fig. 5-2. Processos discriminantes e dispersões discriminantes de dois estímulos

Ora, vê-se pela figura 5-2 que nós podemos utilizar a distribuição R_{j-i} para descobrir a distância entre $\overline{R}_i$ e $\overline{R}_j$, utilizando a dispersão discriminante σ_{j-i}.

Sabemos, pela estatística, que $\sigma_{j-i} = \sqrt{\sigma_i^2 + \sigma_j^2 - 2r_{ij}\sigma_i\sigma_j}$.

De fato, a distância entre R_i e R_j é dada pela fórmula

$$R_j - R_i = z_{ij}\sigma_{ij} \qquad (5.2)$$

$$= z_{ij}\sqrt{\sigma_i^2 + \sigma_j^2 - 2r_{ij}\sigma_i\sigma_j}$$

onde,

R_j e R_i são os processos discriminantes dos estímulos E_j e E_i
σ_i e σ_j são as dispersões discriminantes dos estímulos E_i e E_j
r_{ij} é a correlação entre os pares de processos discriminantes R_i e R_j
z_{ij} é o desvio padrão correspondente à proporção de vezes que E_j foi considerado maior que E_i.

Assim, esta fórmula permite estabelecer as distâncias entre não importa que número de estímulos, resultando daí uma escala de estímulos referentes a algum objeto psicológico que servirão para medir as atitudes dos sujeitos com respeito a este mesmo objeto. Repare que a técnica de Thurstone visa criar um instrumento composto de uma série de estímulos dos quais sabemos a posição escalar de cada um. Ela não escala sujeitos, escala estímulos.

B.2 Procedimentos Experimentais

Thurstone tem várias formas diferentes de coletar a informação para estabelecer os valores escalares dos estímulos, a saber, o método dos intervalos sucessivos, o método dos intervalos aparentemente iguais e o método das comparações emparelhadas. Para exposição de todos estes métodos, recomendamos a obra acima referida de Edwards (1957). Aqui faremos uma breve exposição somente do método das comparações emparelhadas.

No método das comparações emparelhadas são apresentados aos sujeitos uma série de estímulos emparelhados dois a dois. Por exemplo, suponha os seguintes 6 estímulos (frases) sobre Brasília:

1) Viver em Brasília é um prazer
2) Brasília é uma cidade maravilhosa
3) Brasília oferece boas oportunidades de trabalho
4) Brasília dificulta o relacionamento humano
5) É difícil adaptar-se em Brasília
6) É difícil locomover-se em Brasília.

Estes 6 estímulos são apresentados, dois a dois, aos sujeitos, sendo a tarefa destes dizer qual dos dois, na opinião deles, é mais característico de Brasília.

B.3 Procedimentos Analíticos

Ao se fazerem estes julgamentos comparativos entre estímulos, dois a dois, para definir qual deles é 'maior que' em algum atributo (no nosso exemplo, "qual é mais característico de"), produz-se uma tabela de freqüências do tipo $f_{ij} = j > i$ conforme Tabela 5-2, para o nosso exemplo, no qual participaram 100 sujeitos.

Observe que nas diagonais se coloca o N/2, uma vez que o mesmo estímulo nunca é comparado empiricamente consigo mesmo. Se supõe que se ele o fosse, 50% das vezes ele seria considerado maior e 50% das vezes menor que ele mesmo; daí o N/2. A soma das colunas nesta tabela já dá uma idéia da posição escalar de cada item, pois ela diz quantas vezes o estímulo *j* foi considerado maior que todos os outros. Assim, o estímulo de maior posto escalar é o 6, sendo o 3 o de menor posto, resultando na escala de estímulos da seguinte ordem: 3, 1, 2, 4, 5, 6.

Tabela 5-2. Freqüência de vezes em que o *j* é considerado maior que *i* por 100 sujeitos (Matriz F_{ij})

		j					
	Itens	1	2	3	4	5	6
	1	50	60	45	70	80	95
	2	40	50	30	60	50	80
i	3	55	70	50	70	80	90
	4	30	40	30	50	65	85
	5	20	50	20	35	50	60
	6	05	20	10	15	40	50
	Soma	200	290	185	300	365	460

Na diagonal estão os N/2.

Transformando esta matriz de freqüências em uma matriz de proporções, fica mais clara ainda a ordenação dos postos dos estímulos. Para transformar a matriz de freqüências (Tabela 5-2 dos F_{ij}) numa matriz de percentagens ou de proporções (a matriz P_{ij}), divide-se o valor (freqüência) de cada casela pelo número total de respondentes (no caso, N = 100), isto é, $p_{ij} = f_{ij}/N$, o que irá produzir a tabela 5-3.

Tabela 5-3. Matriz P_{ij}.

		j					
	Itens	1	2	3	4	5	6
	1	.500	.600	.450	.700	.800	.950
	2	.400	.500	.300	.600	.500	.800
i	3	.550	.700	.500	.700	.800	.900
	4	.300	.400	.300	.500	.650	.850
	5	.200	.500	.200	.350	.500	.600
	6	.050	.200	.100	.150	.400	.500
	Soma	2.000	2.900	1.850	3.000	3.650	4.600

Com as somas desta matriz 5-3 podemos descobrir os postos dos estímulos numa escala percentílica que vai de 0 a 100, utilizando a fórmula para soma dos totais (Petz & Mayer, 1977), a qual vai dar posições escalares dos estímulos quase idênticos ao método dos *z* (mais complexo) de Thurstone. O método da soma dos totais (ST) consiste em estabelecer os postos escalares dos estímulos pela fórmula

$$(5.3) \quad \overline{R}_j = \frac{S_j - S_1}{A}$$

onde,

$\overline{R}_j$ = posição escalar do estímulo sendo analisado
S_j = soma das proporções da coluna do estímulo sendo analisado
S_1 = soma das proporções do estímulo de menor soma
A = amplitude, isto é, diferença entre a soma maior e a soma menor de proporções.

No nosso exemplo, temos $S_1 = 1{,}850$ e $A = 4{,}600 - 1{,}850 = 2{,}75$. Dando-se ao estímulo com a menor soma o valor 0,0, os demais estímulos terão as seguintes posições escalares

Estímulo	1	2	3	4	5	6
Escala ST	5,5	38,2	0,0	41,8	65,5	100,0

Querendo possibilitar uma escala intervalar de estímulos, Thurstone utilizou a fórmula 5.2 acima apresentada, $R_j - R_i = z_{ij}\sqrt{\sigma_i^2 + \sigma_j^2 - 2r_{ij}\sigma_i\sigma_j}$. Entretanto, esta fórmula tem incógnitas demais; de fato somente o z_{ij} é fornecido pelos dados empíricos. Como está, a fórmula é sem solução. Thurstone, então, fez algumas suposições ulteriores para tornar a fórmula solucionável, das quais surgiram os 5 casos da lei do julgamento comparativo (Edwards, 1957). O caso 5 faz as seguintes suposições:

- não há correlação entre as respostas ao E_i e ao E_j, daí $r_{ij} = 0$
- as dispersões discriminais são iguais, daí $s_i = s_j = s$.

Com estas suposições, a fórmula se reduz a

$$(5.4) \quad R_j - R_i = z_{ij}\sqrt{2\sigma^2}$$

e fazendo $\sqrt{2\sigma^2} = 1$, já que a expressão é uma constante, a fórmula se reduz a

$$(5.5) \quad R_j - R_i = z_{ij}.$$

Com esta fórmula, podemos descobrir as posições escalares dos estímulos, bastando transformar a matriz de proporções (Tabela 5-3) numa matriz de escores z, utilizando a tabela da curva normal, a qual dá as posições escalares dos itens no contínuo do construto, conforme exemplificado na Tabela 5-4.

Tabela 5-4. Matriz Zij.

		i					
	Itens	1	2	3	4	5	6
j	1	,00	,25	-,12	,52	,84	1,65
	2	-,25	,00	-,52	,25	,00	,84
	3	,13	,52	,00	,52	,84	1,28
	4	-,52	-,25	-,52	,00	,39	1,04
	5	-,84	,00	-,84	-,39	,00	,25
	6	-1,65	-,84	-1,28	-1,04	-,25	,00
	Soma (S)	-3,13	-,32	-3,28	-,14	1,82	5,06
	S+3,28	,15	2,96	0	3,14	5,10	8,34

Para inicializar a escala no ponto 0, dá-se ao estímulo com a menor soma dos z o valor 0. No caso, é o estímulo 3, cuja soma é -3,28, a qual se torna 0 se a esta soma somarmos o valor 3,28. Mas fazendo isto com o estímulo 3, devemos

fazer o mesmo com os demais estímulos, como está explicitado na última linha da tabela 5-4; e estes são os valores escalares dos estímulos.

É importante ressaltar que Thurstone oferece testes estatísticos para a verificação da consistência interna da escala resultante, bem como a verificação da adequação das suposições feitas em cada caso utilizado, inclusive o caso 5 (Edwards, 1957).

Utilização da escala de Thurstone. Tendo-se obtido os valores escalares, em termos de desvios padrões, de uma grande série de estímulos, pode-se construir uma escala intervalar, selecionando aqueles (cerca de 20) que se situam a distâncias iguais entre si.

Estes estímulos assim escalonados constituem a escala para a medida das atitudes. Os procedimentos experimentais para aferir as atitudes do sujeito consistem em pedir ao mesmo que escolha o item (estímulo) com o qual ele mais concorda, sendo o valor escalar deste item a medida da atitude do sujeito. Ou, pede-se para o sujeito escolher os três itens com os quais mais concorda e a medida da sua atitude será a média dos valores escalares destes três itens.

A construção de escalas a partir desta lei de Thurstone é extremamente laboriosa. Na verdade, ela se torna quase impossível com um número elevado de itens, dado que a comparação dos mesmos 2 a 2 aumenta geometricamente o número de comparações a serem feitas. Para 10 estímulos temos (10x9)/2 = 45 comparações e para 100 itens temos (100x99)/2 = 4.950. Por isso, Thurstone desenvolveu outras técnicas de construção de escalas de atitude. Uma delas é o método dos intervalos aparentemente iguais (Thurstone & Chave, 1929).

No caso deste método, as afirmações (cerca de 100) sobre um objeto de interesse são impressas em cartões que os sujeitos devem distribuir em 11 pilhas segundo o grau de favorabilidade que, na sua opinião, a afirmação apresenta em relação ao objeto psicológico. As 11 pilhas são erigidas sobre um contínuo de cartões etiquetados de A a K, onde A está ancorado com a expressão 'desfavorável', o K com 'favorável' e o F (o cartão a meio caminho de A e K) com 'neutro'.

O valor escalar dos itens se faz através do cálculo da mediana, tendo como coeficiente de variabilidade o intervalo semi-interquartílico, como na Tabela 5-5.

Tabela 5-5. Cálculo do valor escalar pelo método dos intervalos aparentemente iguais

		Categorias												
Afirmações		A	B	C	D	E	F	G	H	I	J	K	Escala	Q
		1	2	3	4	5	6	7	8	9	10	11		
1	f	2	2	6	2	6	62	64	26	18	8	4		
	p	.01	.01	.03	.01	.03	.31	.32	.13	.09	.04	.02	6.8	1.7
	pa	.01	.02	.05	.06	.09	.40	.72	.85	.94	.98	1.00		
2	f	0	0	0	10	40	28	50	26	28	24	4		
	p	.00	.00	.00	.05	.20	.14	.25	.13	.14	.07	.02	6.9	2.8
	pa	.00	.00	.00	.05	.25	.33	.64	.77	.91	.98	1.00		

f = freqüência; p = proporção; pa = proporção acumulada

Diversas variantes deste método foram propostas (Ballin & Farnsworth, 1941; Seashore & Hevner, 1933; Edwards & Kilpatrick, 1948; Webb, 1951). O próprio Thurstone (Saffir, 1937) apresentou uma variante que chamava de método dos intervalos sucessivos. Para quem quiser se especializar nestes vários tipos de escalas de Thurstone, tornamos a sugerir o livro de Edwards (1957), pois o pouco uso que delas se faz no contexto da Psicometria não justifica maiores detalhamentos neste capítulo.

Note, finalmente, que as escalas de Thurstone não garantem serem unidimensionais. Thurstone considerou que as respostas a todos os estímulos da escala eram dadas em função de uma mesma dimensão, mas não apresenta evidência de tal suposição. Evidentemente, os estímulos alinhados ao longo de uma mesma dimensão não são comparáveis se a escala não for unidimensional. Guttman (em seguida) tentou superar este problema.

C. O Enfoque de Guttman

C.1 Procedimentos Teóricos

Guttman apresentou seu escalograma, para avaliar atitudes, numa série de trabalhos (1944, 1945 1947, 1950).

A parte *teórica* da técnica supõe que a propriedade psicológica possua magnitude e seja unidimensional. Cada item (indicador comportamental) expressa um nível diferente de magnitude, seguindo uma série monotônica crescente (pelo menos de ordem). De sorte que o conjunto de itens da escala expressa o contínuo da propriedade e que, sendo cumulativos, a aceitação de um item de maior nível implica na aceitação de todos os itens inferiores, isto é, de menor posto. A primeira tarefa nesta metodologia, consiste na construção de uma série de itens sobre um construto de tal forma que os itens possam ser escalonados cumulativamente. Esta série consta tipicamente de 6 a 12 itens. Como é efetuada esta tarefa? Guttman diz que é uma questão de intuição e experiência!

C.2 Procedimentos Experimentais

Os procedimentos empíricos consistem em pedir a uma amostra de sujeitos para dizer se concorda ou não com o que o item está afirmando. Se concorda, o item recebe valor 1 e se discorda recebe valor 0. Assim, se os itens estão escalonados cumulativamente, é de esperar que o sujeito que concorda com um item que expressa um certo nível de atitude com respeito ao construto, concordará com todos os itens que têm um nível menor. Desta forma, uma série de itens pode ser escalonada do item mais fraco ou brando até o mais extremo, produzindo uma escala, pelo menos, ordinal. Sendo isto verdade, basta saber o item mais extremo com o qual o sujeito concorda para podermos reproduzir perfeitamente suas respostas nos outros itens. Na realidade, porém, as coisas não acontecem tão certas assim, de sorte que a reprodução das respostas do sujeito nunca será perfeita. Então se pergunta qual é o mínimo de reprodutibilidade das respostas aceitável para se poder dizer que uma escala satisfaz o critério de cumulatividade?

C.3 Procedimentos Analíticos

Com o intuito de verificar a cumulatividade dos itens, Guttman desenvolveu *procedimentos analíticos* para determinar um índice de reprodutibilidade, o qual resulta da comparação entre as respostas correta e incorretamente endossadas. Suponha o seguinte: 4 afirmações sobre um construto psicológico (tendo valor 1 a afirmação mais extrema de atitude) respondidas por 6 sujeitos em termos de estar de acordo (valor 1) ou não-acordo (valor 0). A tabela 5-6 recolhe os resultados fictícios obtidos.

Tabela 5-6. Dados fictícios para a escala de Guttman

Sujeitos	Afirmações (itens)				Soma	Erros
	1	2	3	4		
1	1	1	0	1	3	1
2	0	1	1	1	3	0
3	0	0	1	1	2	0
4	0	0	1	0	1	1
5	0	0	0	1	1	0
Soma	1	2	3	4		

Esta tabela é montada de tal forma que nas colunas estão dispostos os itens em ordem decrescente, do mais extremado ao mais brando, em termos de atitude em relação ao construto e nas linhas estão dispostos os sujeitos, também em ordem decrescente do escore total obtido nos itens (para cada item com o qual está de acordo, o sujeito recebe um ponto). Um item que recebeu o acordo pelo sujeito obtém valor 1 e obtém 0 se o sujeito não o marcou. Assim se forma uma tabela triangular, de tal sorte que acima da diagonal (no canto direito superior) deveriam aparecer somente 1 e abaixo somente 0. No caso em pauta, como o item #1 é o que expressa a atitude mais extrema em relação ao construto, o sujeito que está de acordo com este item deveria necessariamente marcar todos os outros itens, fato que não ocorreu com o sujeito 1 que marcou o item #1 mas não marcou o item #3. Situações desta natureza provocam a ocorrência de 0 acima da diagonal, o que é contado como um erro. Para o cálculo do índice de reprodutibilidade contam-se todos os erros, isto é, os 0 acima da diagonal, que no caso são dois. Assim, o número de valores apropriados na tabela é 20 - 2 = 18. O coeficiente de reprodutibilidade será 18/20 = 0,90. Guttman afirma que o coeficiente deve ser pelo menos de 0,90 para que a escala possa ser considerada adequada.

D. O Enfoque de Likert.

A técnica de Rensis Likert (1932) talvez seja a mais utilizada na construção de escalas psicométricas e é conhecida, desde que Bird (1940) assim a chamou, de método dos pontos somados ("method of summated ratings").

D.1 Procedimentos Teóricos

Em seu pólo teórico, Likert sustenta que uma atitude (propriedade psicológica) constitui uma disposição para a ação. Esta concepção apresentava dificuldades para Likert na época, dado o enfoque do behaviorismo positivista que defendia a atitude como sendo um simples substituto verbal para a ação concreta. Ele defendeu a atitude como um elemento da personalidade, talvez concebido como um construto hipotético, ao afirmar "se de fato tais elementos existem" ("if, in fact, any such elements exist" - 1932: 8). Defendeu igualmente que há uma série de tais construtos de personalidade e não um único; novamente uma diatribe espelhando as disputas da época entre unifatoristas e multifatoristas. Likert nem se pôs a questão da magnitude das propriedades psicológicas (atitude, mais especificamente) pois era para ele uma questão já decidida, isto é, as propriedades psicológicas têm magnitudes, por isso é que podem ser medidas.

A preocupação da escala Likert não consiste em procurar determinar o valor escalar dos itens, como pretendia Thurstone, mas verificar o nível de concordância do sujeito com uma série de afirmações que expressassem algo de favorável ou desfavorável em relação a um objeto psicológico. As afirmações são respondidas numa escala de 3 ou mais pontos, isto é, o sujeito tem que dizer se concorda, está em dúvida ou discorda com o que a frase afirma sobre o objeto psicológico. O número de pontos na escala de resposta varia de 3 a mais de 10, sendo as mais utilizadas as escalas de 5 e 7 pontos. Aliás, o número de pontos utilizados nas escalas Likert parece ser algo irrelevante, como ficou dito no capítulo 3 ponto II. A técnica de Likert consiste em construir uma série (cerca de 20) itens para representar comportamentalmente um construto (para a construção dos itens, vide capítulo 3).

D.2 Procedimentos Experimentais

Os *procedimentos empíricos* consistem em ter os itens respondidos por N sujeitos numa escala de n pontos, como também ficou esclarecido no capítulo 3.

D.3 Procedimentos Analíticos

Os *procedimentos analíticos* visam determinar a seleção final dos itens e a avaliação dos parâmetros psicométricos da escala.

Likert utilizou duas técnicas para selecionar os itens: 1) análise da consistência interna dos itens (o teste *t* de Student) e 2) a correlação. A análise da correlação consistia em correlacionar cada item com o restante dos itens: se esta correlação não fosse significativa, o item era descartado porque não estava medindo a mesma coisa que o restante dos itens. Como esta técnica era laboriosa, pois ainda não havia a disponibilidade dos computadores, Likert sugeriu utilizar a análise da consistência interna em termos do seu poder de discriminação de grupos-critério, formados estes à base do escore total que os sujeitos obtêm na escala. Assim, um teste *t* entre as médias de cada item, obtidas pelo grupo superior e inferior (os 30% escores superiores e 30% inferiores na escala) definem a discriminabilidade dos itens. En-

tretanto, as análises mais modernas da TRI parecem mais promissoras neste particular, pois elas oferecem até três parâmetros para os itens; discriminação, dificuldade e resposta ao acaso.

Na análise da própria escala, importa em verificar a validade e a precisão. Uma análise importante da escala consiste em verificar a unidimensionalidade suposta da mesma. Tipicamente se utiliza para tanto a análise fatorial, pois o modelo de Likert não garante que os itens estejam medindo a mesma coisa. A análise da fidedignidade é comumente feita através da análise da consistência interna dos itens através do coeficiente alfa de Cronbach. Mas qualquer das técnicas de validade e precisão podem ser aqui utilizadas.

Fica ainda em dúvida se a escala de Likert produz medidas somente ordinais ou se chegam a ser de intervalo. Na verdade, com os dados empíricos coletados com a escala, pode-se avaliar o valor escalar das categorias utilizadas (os pontos) na escala de resposta (Edwards, 1957) e, a partir daí, utilizar estes valores escalares para as categorias. Tal procedimento, contudo, tira a leveza e a facilidade de trabalhar com as escalas tipo Likert. Edwards e Kenny (1946), aliás, verificaram que escalas construídas no estilo Likert (considerando as categorias 1, 2, 3 etc. como intervalos iguais) correlacionam em torno de 0,90 com escalas de intervalos aparentemente iguais de Thurstone. Concluem, ainda, que dada a facilidade de construção e utilização, as escalas tipo Likert se apresentam com grande vantagem sobre as de tipo Thurstone. Outros autores (Nunnally, 1978; Alwin, 1973; Sewell, 1941; McIver & Carmines, 1981; Greene & Carmines, 1979; Zeller & Carmines, 1980; Carmines & Zeller, 1980) também acham que usar pesos para a escala de resposta não compensa o esforço e o ganho é irrisório.

Finalmente, a interpretação dos escores numa escala Likert não é imediatamente aparente. De fato, o que significa receber um escore de 30 numa escala de 40 itens? A maneira mais apropriada (Edwards, 1957; McIver & Carmines, 1981) de interpretar os escores da escala Likert consiste em posicioná-los relativamente ao grupo que respondeu a escala, isto é, criar normas baseadas no grupo de resposta, onde a média do grupo será o ponto de referência. O próprio Likert já alertava para este fenômeno, quando dizia que, como no caso dos testes de inteligência, também os escores de atitude devem ser expressos em termos de normas baseadas no grupo respondente (Likert, 1974). Desta forma, os escores da escala Likert são expressos em escores padrões (z), os quais indicam quanto um dado sujeito se afasta da média. Assim, um sujeito que recebeu o escore 30 numa escala de 40 itens, cuja média no grupo foi de 20 e com um desvio padrão de 5, terá um escore padrão z de 2, isto é, (20 - 30) / 5. Então o escore 30 se situa a 2z acima da média, ou seja, este escore põe o sujeito no percentil 98, o que representa uma atitude extremamente favorável no atributo medido.

E. Escala Multidimensional

E.1 Procedimentos Teóricos

As escalas de tipo Thurstone, Guttman e Likert são ditas unidimensionais porque elas visam avaliar os sujeitos em apenas um traço psicológico. Falando-se de

uma escala multifatorial, neste caso, entender-se-ia um conjunto de várias escalas, cada uma medindo um fator ou traço independentemente. Entretanto, um objeto psicológico pode ser avaliado sob vários aspectos ou traços simultaneamente. Por exemplo: um candidato à presidência pode ser avaliado em termos de sua filiação partidária (liberal vs. conservador), recebendo uma pontuação nesta escala; ao mesmo tempo e independentemente, ele pode ser avaliado em termos de sua juventude (jovem vs. velho), recebendo nesta escala uma outra avaliação, independente da que recebeu na primeira escala. Teríamos aqui, então, duas escalas unidimensionais, produzindo dois escores independentes. Contudo, pode-se pedir uma avaliação simultânea do candidato em termos de ambos os atributos, a saber, filiação partidária e juventude. No primeiro caso, o candidato teria dois escores: um em filiação partidária e outro em juventude. No segundo caso, entretanto, o candidato receberia apenas um escore, mas definido em termos de duas dimensões, que seria ilustrado num espaço bidimensional, onde um ponto é expresso por duas coordenadas. Assim, o escore dele neste caso seria expresso como X_{ij} e não por X_i e X_j. Continuando nesta ilustração, o mesmo candidato poderia ser avaliado numa série de *n* traços simultaneamente, de sorte que o escore dele poderia ser expresso num espaço n-dimensional, com tantos subscritos quantos os traços sob os quais ele foi avaliado. Estas são as escalas multidimensionais. No caso de uma avaliação em termos de duas dimensões, o escore do candidato poderia cair em qualquer um dos quatro quadrantes que resultam do espaço bidimensional, como na Figura 5-3, onde o candidato X_1 se situa em (-1,2).

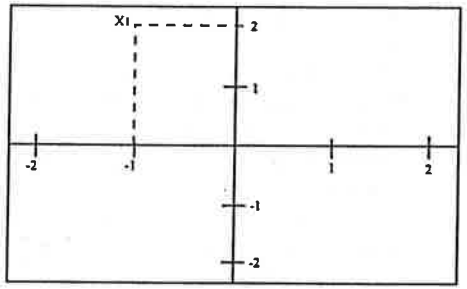


Fig.5-3. Sujeito X_1 situado em um espaço bidimensional

Ao ser avaliada uma série de candidatos, cada um deles seria expresso por um ponto que se situaria num dos quadrantes. Os pontos mantêm uma relação de proximidade (distância) entre si expressa simultaneamente com respeito a dois traços: filiação partidária (eixo horizontal) e juventude (eixo vertical). Para *n* dimensões, o ponto de cada candidato teria, obviamente, proximidades entre si com respeito a *n* eixos (num espaço n-dimensional).

E.2 Procedimentos Experimentais

A técnica para levantar os dados de escalas multidimensionais consiste em pedir ao(s) sujeito(s) para avaliar(em) um objeto psicológico (candidato), não em um traço de cada vez, mas em comparar vários objetos psicológicos em vários

traços. Por exemplo: Dados os candidatos A B C D, avaliar se os candidatos A e B são mais semelhantes (próximos, iguais, etc.) entre si que os candidatos C e D. Assim, a técnica para a coleta da informação usa termos que se referem a "distância psicológica" ou "proximidade psicológica". Esta proximidade vem designada sob várias expressões, tais como parentesco, dependência, associação, complementaridade, substitutividade, distância, proximidade, interação, etc.

Um exemplo poderá ilustrar os procedimentos da técnica das escalas multidimensionais. Suponha quatro candidatos à presidência (A, B, C, D). Os respondentes reagem à instrução de emparelhar 2 a 2 os candidatos e dizer qual é o candidato preferido entre os dois (i > j). Deste procedimento podem surgir os seguintes dados:

		I			
	Candidato	A	B	C	D
	A	-			
J	B	7	-		
	C	5	8	-	
	D	3	6	9	-

O candidato A foi preferido 7 vezes a B, 5 vezes a C e 3 vezes a D, etc. Estes números podem ser considerados como indicando distâncias entre os candidatos e serem expressos numa matriz de distâncias. Neste caso, surge uma matriz simétrica, onde a distância d_{ij} é igual a d_{ji} e tendo o valor 0 na diagonal, como segue (note, entretanto, que nem sempre d_{ij} deve ser necessariamente igual a d_{ji}):

$$\begin{bmatrix} 0 & 7 & 5 & 3 \\ 7 & 0 & 8 & 6 \\ 5 & 8 & 0 & 9 \\ 3 & 6 & 9 & 0 \end{bmatrix}$$

E.3 Procedimentos Analíticos

Com referência ao *pólo analítico*, a fórmula para cálculo das distâncias é a fórmula euclidiana normalmente utilizada para distâncias, qual seja:

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + \ldots (x_{ir} - x_{jr})^2}$$

ou sumariamente

$$d_{ij} = \sqrt{\sum_{r=1}^{R} (X_{ir} - X_{jr})^2}$$

Onde *r* corresponde ao número de dimensões (fatores) sob as quais os candidatos foram avaliados, no nosso caso sendo supostamente duas (filiação partidária e juventude).

Para a determinação da dimensionalidade que subjaze às proximidades encontradas entre os objetos psicológicos avaliados (candidatos, no nosso caso), há vários enfoques estatísticos, tanto paramétricos quanto não-paramétricos: "Coombs' unfolding technique" (Coombs, 1964), o modelo de Tucker e Messick (Tucker & Messick, 1963), o modelo de Torgerson (1958), o modelo de Ekman (1963, 1965), etc. Para tanto, consultem-se Kruskal e Wish (1991) e Delbeke (1968). Existem, igualmente, uma série de programas de computador para as análises com escalas multidimensionais, tais como o Multilog da Scientific Software, Inc. (1991) e outros (Kruskal & Wish, 1991: 79).

A tecnologia das escalas multidimensionais tem sido usada por psicólogos, sociólogos, antropólogos, economistas e educadores (Uslaner, apud Kruskal & Wish, 1991). Seu uso em Psicologia, no entanto, não tem sido muito extenso, apesar do seu caráter promissor na determinação da dimensionalidade nas preferências psicológicas dos indivíduos. O caráter de complexidade estatística talvez seja uma das razões para o pouco uso que se faz das escalas multidimensionais.

Conclusão

Apesar dos muitos problemas que ainda existem na teoria da medida em ciências sociais e do comportamento, o uso de escalas, especialmente em Psicologia Social e da Personalidade, além de apresentar uma história de mais de meio século, é ainda muito difundido. Esta ocorrência não pode ser considerada fortuita, mas deve proceder do fato de que as medidas escalares são capazes de produzir conhecimento válido nas ciências do comportamento. As várias técnicas apresentadas (Likert, Thurstone, Guttman, etc.) têm apresentado razoável consistência, tanto em sua estrutura interna quanto nos resultados obtidos através delas. Todas essas técnicas, na verdade, oferecem procedimentos estatísticos que permitem avaliar essa consistência interna. Quanto à consistência dos resultados que produzem, a situação das escalas existentes e as próprias técnicas propostas para a sua construção não aparecem ainda como empolgantes. É possível, e quiçá provável, que este fenômeno se deva em grande parte à falta de definição mais precisa destas mesmas técnicas quanto aos procedimentos teóricos envolvidos na elaboração dos instrumentos. Há uma preocupação grande, e louvável, referente à adequação dos procedimentos estatísticos, mas estes não dão dicas fundamentais quanto ao verdadeiro problema da escala, que é a construção de um instrumento válido, isto é, que de fato esteja medindo algo de psicologicamente relevante. Sem uma boa teoria psicológica que a fundamente, a escala pode até aparecer estatisticamente perfeita e consistente, mas medindo nada de relevante ou medindo algo desconhecido.

Bibliografia

Alwin, D.F. (1973). The use of factor analysis in the construction of linear composites in social research. *Sociological Methods and Research, 2*, 191-214.

Ballin, M. & Farnsworth, P.R. (1941). A graphic rating method for determining the scale values of statements in measuring social attitudes. *Journal of Social Psychology, 13*, 323-327.

Bendig, A.W. (1954). Reliability and the number of rating scale categories. *Journal of Applied Psychology, 38*, 38-40.

Bird, C. (1940). *Social psychology*. New York: Appleton-Century-Crofts, Inc.

Carmines, E.G. & Zeller, R.A. 1980). *Reliability and validity assessment*. Beverly Hills, CA: SAGE.

Clausen, A.R. & Van Horn, C.E. (1977). How to analyze too many rolls calls and related issues in dimensional analysis. *Political Methodology, 4*, 313-331.

Coombs, C.H. (1964). *A theory of data*. New York: John Wiley.

Delbeke, L. (1968). *Construction of preference spaces*. Louvain, Belgium: Publications of the University of Louvain.

Edwards, A.L. & Kilpatrick, F.P. (1948). A technique for the construction of attitude scales. *Journal of Applied Psychology, 32*, 374-384.

Edwards, A.L (1957). *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts, Inc.

Edwards, A.L. & Kenny, K.C. (1946). A comparison of the Thurstone and Likert techniques of attitude scale construction. *Journal of Applied Psychology, 53*, 72-83.

Ekman, G., & Künnapas, T. (1963). A further of direct and indirect scaling methods. *Scandinavian Journal of Psychology, 4*, 77-80.

Ekman, G. & Sjöberg, L. (1965). Scaling. *Annual Review of Psychology, 16*, 451-474.

Goldsamt, M.R. (1971). Effects of scoring method and rating scale length in extreme response style measurement. Unpublished doctoral dissertation, University of Maryland.

Greene, V.L. & Carmines, E.G. (1979). Assessing the reliability of linear composites. In K.F. Shuessler (Ed.), *Sociological methodology*. San Francisco, CA: Jossey-Bass, 160-175.

Guilford, J.P. (1954). *Psychometric methods*, 2d ed. New York: McGraw-Hill.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139-150.

Guttman, L. (1945). *Questions and answers about scale analysis*. Research Branch, Information and Education Division, Army Service Forces. Report D-2.

Guttman, L. (1947). On Festinger's evaluation of scale analysis. *Psychological Bulletin, 44*, 451-465.

Guttman, L. (1947). The Cornell technique for scale and intensity analysis. *Educational and Psychological Measurement, 7*, 247-280.

Guttman, L. (1950). The problem of attitude and opinion measurement. In S.A. Stouffer et al., *Measurement and Prediction*. Princeton, N.J.: Princeton University Press, 46-59.

Guttman, L. (1950). The basis for scalogram analysis. In S.A. Stouffer et al., *Measurement and Prediction*. Princeton, N.J.: Princeton University Press, 60-90.

Jenkins, J.J., Russell, W.A., & Suci, G.J. (1957). An atlas of semantic profiles for 360 words. In *Studies on the role of language in behavior*. Tech. Rep. No. 15. Minneapolis: University of Minnesota.

Jones, R.R. (1968). Differences in response consistency and subject's preferences for three personality inventory response formats. *Proceedings of the 67th Annual Convention of the American Psychological Association, 3*,, 247-248.

Komorita, S.S. (1963). Attitude content, intensity, and the neutral point on a Likert scale. *Journal of Social Psychology, 61*, 327-334.

Kruskal, J.B. & Wish, M. (1991). *Multidimensional scaling*. Newbury Park, CA: Sage Publications.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology , 140*, 55 ps.

Likert, R. (1974). A method of constructing an attitude scale. In G.M. Maranell (Ed.), *Scaling: A sourcebook for behavioral scientists*. Chicago: Aldine, 233-243.

Luce, R.D. (1977). Thurstone's discriminal processes fifty years later. *Psychometrika, 42*(4), 461-489.

Matell, M.S. & Jacoby, J. (1972). Is there an optimal number of alternatives for Likert-scale items? *Journal of Applied Psychology, 56(6)*, 506-509.

Matell, M.S. & Jacoby, J. (1971). Is there an optimal number of Likert scale items? Study I: Reliability and validity. *Educational and Psychological Measurement, 31*, 657-674.

McIver, J.P. & Carmines, E.G. (1981). *Unidimensional scaling*. Newbury Park, CA: SAGE.

Nunnally, J.C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Petz, B. & Mayer, D. (1977). Comparison between Thurstone's "Law of comparative judgments" scale and the "Sum of totals" scale. *Acta Instituti Psychologici Universitatis Zagrabiensis*, Zagreb, 55-64.

Saffir, M.A. (1937). A comparative study of scales constructed by three psycho-physical methods. *Psychometrika, 2*, 179-198.

Scientific Software (1991). *Multilog User's Guide*, Version 6.0. Cicago, IL: Scientific Software, Inc.

Seashore, R.H. & Hevner, K. (1933). A time-saving device for the construction of attitude scales. *Journal of Social Psychology, 4*, 366-372.

Sewell, W.H. (1941). The development of a sociometric scale. *Sociometry, 5*, 279-297.

Stevens, S.S. (1951). Mathematics, measurement, and psychophysics. In S.S. Stevens (ed.), *Handbook of experimental psychology*. New York: John Wiley & Sons, Inc., 1-49.

Thurstone, L.L (1927). A law of comparative judgment. *Psychological Review, 34*, 273-286.

Thurstone, L.L (1927). Psychophysical analysis. *American Journal of Psychology, 38*, 368-389.

Thurstone, L.L (1927). The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology, 21*, 384-400.

Thurstone, L.L (1927). Equally often noticed differences. *Journal of Educational Psychology, 18*, 289-293.

Thurstone. L.L. & Chave, E.J. (1929). *The measurement of attitude*. Chicago, Ill.: University of Chicago Press.

Torgerson, W.S. (1958). *Theory and methods of scaling*. New York: John Wiley.

Tucker, L.R. & Messick, S. (1963). An individual differences model for multidimensional scaling. *Psychometrika, 28*, 333-367.

Van der Veer, F., Howard, K.I., & Austria, A.M. (1970). Stability and equivalence scores based on three different response formats. *Proceedings of the 78th Annual Convention of the American Psychological Association, 5*, 99-100.

Webb, S.C. (1951). A generalized scale for measuring interest in science subjects. *Educational and Psychological Measurement, 11*, 456-469.

Zeller, R.A. & Carmines, E.G. (1980). *Measurement in the social sciences: The link between theory and data*. New York: Cambridge University Press.

O Diferencial Semântico

Luiz Pasquali

Osgood (Osgood & Suci, 1952; Osgood, Suci, & Tannenbaum, 1957) desenvolveu o pólo teórico de sua posição e uma técnica para medir o conceito de significado ("meaning"), a qual chamou de Diferencial Semântico.

A. Procedimentos Teóricos

A.1 - O Significado

O significado representa um estado cognitivo, entendido como um processo de mediação representativa da realidade. O processo de mediação é concebido como algo que se intercala entre o estímulo e a resposta, como segue:

$$E \rightarrow r \rightarrow s \rightarrow R$$

O processo mediativo é constituído pela relação r→s, significando que o *E* (estímulo externo) detona um processo psicológico interno composto de uma reação interna (r) ao *E* externo, reação que Osgood, seguindo Morris (1946), chama de "disposição", a qual provoca estimulações internas (s), levando este processo s→r a uma ação externa (R). Este processo cognitivo dá o significado ao *E* para produzir a ação *R*. Através de estudos fatoriais, Osgood chegou a definir este processo cognitivo mediativo como sendo caracterizado por várias dimensões, sendo particularmente importantes três grandes fatores: o processo apresenta um aspecto avaliatório (emocional - "evaluative"), um de poder ("potency") e um de atividade ("activity"). Isto quer dizer que o significado das coisas, dos conceitos, etc. para o sujeito varia em função destes três fatores.

A.2 - O Espaço Semântico

Esses fatores ou dimensões podem ser visualizados como vetores de um espaço, chamado espaço semântico, porque representa o espaço do significado. Este espaço é multidimensional e os conceitos podem ser expressos pelas suas coordenadas neste espaço. Assim, se o significado se situa num espaço de três dimen-

sões, composto pelos três fatores acima referidos, então o significado de pai, por exemplo, seria expresso por um ponto neste espaço tridimensional como (x,y,z), expressando a convergência dos três vetores de avaliação, poder e atividade. Osgood ainda assume que estas dimensões do espaço semântico são ortogonais, isto é, são independentes umas das outras. Esta suposição é feita para tornar a interpretação do significado mais eficiente e econômica. De fato, este espaço semântico pode apresentar infinitas dimensões, mas a grande parte delas são colineares, isto é, vão na mesma direção; então todas as que vão na mesma direção não acrescentam nada de substancialmente novo para compreender o significado. Assim, a tarefa para definir este espaço semântico de modo eficiente consiste em determinar o número mínimo necessário de dimensões ortogonais ou eixos que exaurem a dimensionalidade deste espaço. Evidentemente, a técnica lógica para efetuar esta tarefa é a análise fatorial, que precisamente visa estabelecer tais eixos ortogonais. Note, entretanto, que a suposição de ortogonalidade não é necessária para trabalhar com o conceito de espaço semântico de Osgood, bem como com a sua técnica do Diferencial Semântico, pois se pode trabalhar perfeitamente e, quiçá, com vantagem com a rotação oblíqua. O próprio Osgood e outros (1957) afirmam que o espaço comportará tantas dimensões quantas se puder identificar e medir de uma forma fidedigna. Obviamente, quanto mais independentes entre si estas dimensões, mais fácil será a sua interpretação.

Os vetores do espaço semântico apresentam duas propriedades, a saber, direção e distância. Considerando que o espaço tem uma origem, então a direção do vetor diz respeito à orientação que ele assume com relação a esta origem e aos demais vetores; a distância corresponde à intensidade da reação dos sujeito com respeito ao estímulo e é expressa pelo afastamento do ponto final do vetor com relação à origem, como ilustra a figura 6-1.

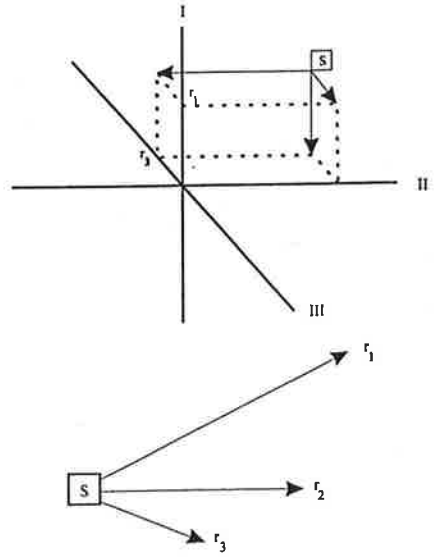


Fig. 6-1. Relação entre mediação (r) e espaço semântico

A.3 - O Diferencial Semântico

1 - Conceituação

Como o processo mediativo é o que determina o matiz da ação do sujeito (a resposta *R*), é de importante utilidade se poder medi-lo. É o que pretende fazer a técnica do Diferencial Semântico (DS). O objetivo do Diferencial Semântico consiste em especificar as condições de estímulo e de resposta sob as quais este processo funciona. Ela consiste em se criar uma série de escalas bipolares (itens) para descrever um conceito. Para ser representativa, esta série deve cobrir as dimensões do significado acima assinaladas (os três fatores). Estas escalas são apresentadas em forma de adjetivos descritivos bipolares. O formato das escalas ou adjetivos bipolares são apresentados como segue:

PAI

forte ___:___:___:___:___:___:___ fraco
bom ___:___:___:___:___:___:___ mau
ativo ___:___:___:___:___:___:___ passivo.

2 - Construção de um Diferencial Semântico

A elaboração de um Diferencial Semântico consiste na escolha de 1) os conceitos que servirão de estímulo aos quais os sujeitos deverão dar uma resposta e 2) as escalas que servirão de itens que os sujeitos irão responder com respeito a estes conceitos.

A escolha dos conceitos é questão de interesse do pesquisador. Talvez a única dica que se pode aqui dar é de que os conceitos sejam relevantes para o pesquisador e que sejam unívocos, isto é, que não sejam conceitos confusos e ambíguos.

A escolha das escalas já é mais estruturada. De fato, como as escalas devem ser o mais representativas possível dos conceitos escolhidos, bem como dos fatores descobertos pela pesquisa do Diferencial Semântico (avaliação, potência, atividade e outros), elas deverão ser 1) relevantes aos conceitos (relevância) e 2) ter alta carga fatorial nos fatores (composição fatorial).

Osgood, Suci e Tannenbaum (1957) dão um elenco de itens bipolares já utilizados em muitas pesquisas e que podem ser úteis na hora que se vai construir um diferencial semântico, como aparece na tabela 6-1. Eles dão, inclusive, a carga fatorial destas escalas. Trabalho similar foi feito no Brasil por Alves Pereira (1986).

Escalas para alguns outros fatores:

- *Estabilidade*: sóbrio - bêbado, estável - mutável, racional - intuitivo, sadio - insano, cauteloso - apressado, ortodoxo - herético;
- *Tensidade*: angular - redondo, reto - curvo, afiado - cego;
- *Novidade*: novo - velho, inusitado - usual, jovem - maduro;
- *Receptividade*: saboroso - sem gosto; renovado - cansado; colorido - sem cor, interessante - cansativo, picante - suave, sensitivo - insensível;

- *Agressividade*: agressivo - defensivo, decorado - simples, perto - longe, heterogêneo - homogêneo, tangível - intangível, inerente - estranho, molhado - seco, simétrico - assimétrico, competitivo - cooperativo, formatado - desformado, periódico - errático, sofisticado - ingênuo, público - privado, humilde - orgulhoso, objetivo - subjetivo, econômico - generoso.

Tabela 6-1. Escalas para os fatores avaliação, potência e atividade

Fatores					
Avaliação		Potência		Atividade	
bom	mau	duro	mole	ativo	passivo
otimista	pessimista	forte	fraco	excitado	calmo
completo	incompleto	severo	leniente	quente	frio
oportuno	inoportuno	tenaz	dócil	intencional	involuntário
altruísta	egoísta	constrangido	livre	rápido	lento
sociável	insociável	apertado	espaçoso	complexo	simples
meigo	cruel	pesado	leve	calmo	nervoso
agradecido	ingrato	sério	humorístico	silencioso	barulhento
harmonioso	dissonante	opaco	transparente	pacífico	violento
limpo	sujo	grande	pequeno	seguro	perigoso
claro	escuro	masculino	feminino	fácil	difícil
gracioso	desajeitado	grande	pequeno	construtor	destruidor
prazeroso	doloroso	comprido	curto	natural	artificial
bonito	feio	muito	pouco	vivo	morto
exitoso	fracassado	largo	estreito	livre	preso
alto	baixo	total	parcial	justo	injusto
significativo	sem sentido	alto	baixo	justo	injusto
importante	insignificante				
progressivo	regressivo				
verdadeiro	falso				
positivo	negativo				
honrado	desonrado				
crente	cético				
sábio	tolo				
sadio	doentio				
útil	inútil				
magnífico	horrível				
agradável	desagradável				
amigo	inimigo				
gostoso	ruim				

B - Procedimentos Experimentais

Um Diferencial Semântico é tipicamente composto de um ou mais conceitos que são avaliados contra uma série de escalas bipolares (entre 10 e 20)

B1 - Formato das escalas

Osgood experimentou duas formas de escalas; numa Forma I os conceitos eram diferentes para cada nova escala e noutra (Forma II) toda a série de escalas era avaliada contra um conceito de cada vez, como mostra a figura 6-2.

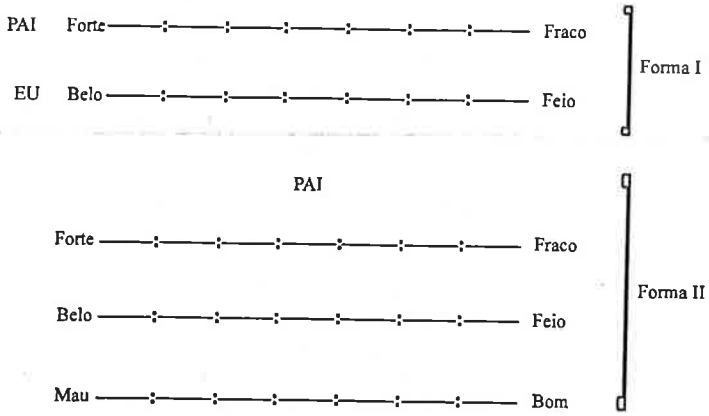


Fig. 6-2. As duas formas do formato das escalas do Diferencial Semântico

A forma I foi prevista para evitar o efeito de halo, já que o sujeito devia mudar de conceito a cada escala avaliada, não facilitando assim que ele comparasse seus julgamentos novos com os anteriores. Entretanto, esta escala possibilitava que o próprio conceito (o significado do conceito) sendo avaliado mudasse de significado de uma escala para a outra. Por isso, Osgood preferiu a forma II por manter o conceito constante e ter se mostrado mais apreciada pelos respondentes.

B2 - Instruções

Segundo Osgood, Suci e Tannenbaum (1957), as instruções devem orientar o respondente com referência à natureza da tarefa, o significado das posições nas escalas e a atitude que ele deve assumir diante da tarefa. Instruções típicas seriam como as seguintes:

O objetivo deste teste consiste em medir o *significado* que as coisas têm para diferentes pessoas, avaliando-as contra uma série de escalas. Ao responder, faça seus julgamentos com base no que estas coisas significam *para você*. Neste folheto você vai encontrar em cada página um conceito diferente, seguido de uma série de escalas bipolares, isto é, ancoradas nos extremos por adjetivos opostos. Sua tarefa consiste em avaliar o conceito, marcando em cada escala o pólo que, segundo você, melhor descreve o tal conceito.

Exemplo: Seja o conceito "amigo" e as escalas as seguintes

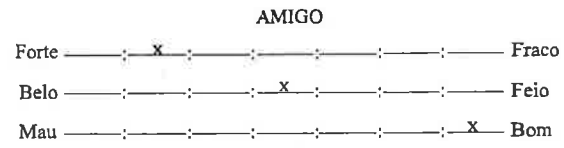


Fig. 6-3. Escalas para avaliar o significado de "amigo"

Se você marcou as três escalas como no exemplo, significa que você acha o AMIGO *bastante forte, nem belo nem feio* e *muito bom*. Assim, quanto mais perto de um dos pólos da escala você marcar, mais característico é do conceito o adjetivo que ancora este pólo.

Faça a marca no meio dos espaços, não em cima das divisões. Não reflita demais, pois o primeiro pensamento é sempre o melhor. Responda todas as escalas.

C - Procedimentos Analíticos

Aplicando um Diferencial Semântico, por exemplo, com 10 escalas com referência a 3 conceitos a uma amostra de 100 sujeitos, teremos uma matriz cúbica de 10 x 3 x 100 caselas. Agora, o interesse do pesquisador pode ser variado. Ele pode estar interessado no perfil dos três conceitos para cada sujeito; neste caso vai ter 100 perfis do tipo da figura 6-4. Normalmente, o pesquisador estará interessado no perfil dos conceitos em grupos de sujeitos; neste caso o perfil será o mesmo da figura 6-4, apenas que será para um grupo e não para um sujeito individual. Entretanto, um perfil de conceitos sobre todas as escalas normalmente não é interessante e ilustrativo, dado que várias escalas estão avaliando características semelhantes. Assim, faz mais sentido primeiramente fazer uma análise fatorial das escalas, tomando todos os dados, combinando sujeitos e conceitos, para descobrir que fatores elas estão medindo e fazer, em seguida, o perfil dos conceitos em cima dos fatores. Neste caso é preciso criar os escores fatoriais e não mais trabalhar com os escores obtidos em cada escala individualmente. A seguir, falaremos desta última maneira de analisar o Diferencial Semântico.

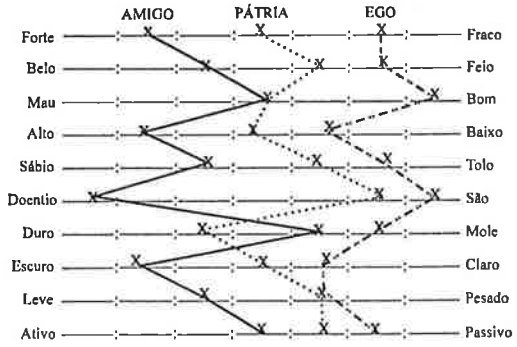


Fig. 6-4. Perfil de 3 conceitos em 10 escalas

Para se poder efetuar as análises do Diferencial Semântico em seguida propostas, é preciso se transformar a escala gráfica de resposta em uma escala numérica. De fato, as sete posições desta escala podem ser expressas de duas formas:

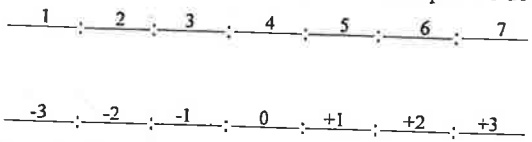


Fig. 6-5. Transformações da escala do DS

Para fins de análise, ambas são idênticas. A segunda tem uma vantagem intuitiva de colocar o ponto neutro em 0 em vez de 4 como faz a primeira. Provavelmente, é mais simples e cômodo trabalhar com a primeira forma, sobretudo para a digitação e crítica dos dados.

Os *procedimentos analíticos* consistem em verificar, além da composição fatorial do Diferencial Semântico, pelo menos, dois aspectos: o significado que o conceito tem para o sujeito ou grupo de sujeitos em termos dos fatores que a análise fatorial detectar no instrumento e a distância que diferentes conceitos mantêm entre si para estes mesmos sujeitos e, ainda, a estrutura semântica que vários conceitos mantém entre si.

C1 - Composição Fatorial das Escalas

Antes de proceder a qualquer análise dos dados, é necessário verificar quantos e quais fatores as escalas utilizadas no Diferencial Semântico estão representando. Uma análise fatorial se faz necessária para decidir tal questão. Talvez pareça estranha esta exigência, uma vez que Osgood, Suci e Tannenbaum (1957) já descobriram que escalas estão medindo que fatores. Isto é verdade, mas para os USA. Quem me garante que aquelas mesmas escalas estão medindo os mesmos fatores no Brasil e, sobretudo, no contexto da minha pesquisa de agora? É preciso verificar empiricamente, isto é, é preciso validar o Diferencial Semântico para o Brasil e, mesmo, para cada pesquisa que faço, pois cada vez que ele é modificado, a sua validade deve ser verificada. Do contrário, estamos laborando em puras hipóteses, quando a pesquisa é para testar empiricamente estas hipóteses. No Brasil, Alves Pereira (1986) já realizou um estudo razoável do Diferencial Semântico.

Esta análise fatorial é das escalas. Por isso, é preciso somar sobre conceitos e sobre sujeitos. Isto é, se há três conceitos, para cada escala temos três respostas do mesmo sujeito: é como se fossem três sujeitos diferentes. Claro, você pode fazer a análise fatorial das escalas dentro de cada conceito, somando apenas sobre os sujeitos. Mas neste caso, como será possível comparar os conceitos se os fatores que surgirem forem diferentes de conceito para conceito? Nós queremos uma estrutura ou composição fatorial das escalas que se aplique a todos os conceitos da pesquisa; daí porque se soma sobre conceitos e sobre sujeitos. Lembre-se, contudo, como dissemos no capítulo 3, que para se fazer análise fatorial adequada são exigidas amostras grandes de sujeitos (10 sujeitos por escala ou 100 sujeitos por fator).

C2 - Significado de um Conceito

Somando-se os escores de cada item de um fator e dividindo pelo número de itens no fator, obtém-se o escore fatorial do sujeito no conceito para aquele fator. Lembre-se que para somar os escores das escalas, estas devem estar unidirecionadas, isto é, todas as escalas devem ter o pólo positivo de um mesmo lado da escala de resposta; se não for o caso, a escala de resposta de -3 a +3 deve ser invertida antes de somar seu escore aos escores do restante das escalas.

Assim, havendo três fatores, o sujeito terá três escores fatoriais que representam semanticamente o conceito. Desta forma, se um sujeito avaliou cinco conceitos em nove escalas (três para cada fator), podemos ter os dados da Tabela 6-2 (dados fictícios).

Tabela 6-2. Matriz de escores em cinco conceitos avaliados por um sujeito em 9 escalas

Escalas		Conceitos				
		Pai	Herói	Destino	Guerra	Paz
Bom	Ruim	3	3	0	-3	3
Doce	Amargo	2	3	0	-3	2
Agradável	Desagradável	2	2	-1	-3	3
Forte	Fraco	1	3	3	2	3
Grande	Pequeno	1	3	-2	3	1
Poderoso	Impotente	2	2	3	3	0
Ativo	Passivo	1	2	2	-3	1
Rápido	Vagaroso	1	3	-2	2	1
Cortante	Embotado	0	0	1	2	1

Sendo que os primeiros três itens medem o fator avaliação, os três do meio o fator de potência e os três últimos o fator atividade, os escores fatoriais deste sujeito serão os da tabela 6-3.

Tabela 6-3. Escores fatoriais de um sujeito em cinco conceitos

Escalas	Conceitos				
	Pai	Herói	Destino	Guerra	Paz
Avaliação	2,33	2,67	-0,33	-3,00	2,67
Potência	1,33	2,67	1,33	2,67	1,33
Atividade	0,67	1,67	0,33	0,33	1,00

Assim, este sujeito considera o pai como muito bom, bastante poderoso e mais ou menos ativo, ao passo que considera guerra como sendo péssima, muito poderosa e mais ou menos ativa, etc.

Com estes dados fatoriais não há muito mais o que se dizer sobre o significado que o sujeito atribui a estes conceitos. Mais interessante é se estes conceitos puderem ser comparados entre si e, mesmo, com os significados dados aos mesmos conceitos por outros sujeitos ou grupos de sujeitos. É o que pretende fazer a fórmula da distância, que veremos a seguir.

C3 - Similaridade e Diferença no Significado (Distância)

1 - O Espaço Semântico

No Diferencial Semântico, o conceito está situado num espaço semântico, geralmente de três dimensões (avaliação, potência, atividade). Seu ponto de fixação neste espaço é definido pelas coordenadas dele com as *n* dimensões deste espaço, dimensões geralmente expressas em eixos ortogonais. Este espaço tem sua origem no ponto 0 e sua extensão máxima vai de -3 a +3 (na forma de escala -3 a +3). Assim, os dados da tabela 6-3 geram três eixos para cada conceito, indicando que os conceitos estão situados num espaço tridimensional. O caso do conceito pai, por exemplo, tem eixos ortogonais que se encontram no ponto (2,33; 1,33; 0,67), que é o ponto onde se situa o conceito pai no espaço semântico tridimensional. Cada um dos outros conceitos tem seu ponto no espaço definido pelos respectivos eixos, que são os valores expressos na tabela.

2 - A Distância Semântica

Tendo estes cinco conceitos da tabela 6-3 ancorados no espaço semântico pelas coordenadas expressas pelos valores fatoriais da tabela, pode-se perguntar "qual é a distância entre eles?" Ou, "quais deles são semanticamente mais similares ou mais diferentes?"

A resposta a esta pergunta é dada pelo conceito de distância. Note preliminarmente que a correlação, que avalia a covariância entre os conceitos, não é capaz de responder à questão. Veja, por exemplo, os escores em quatro fatores dos seguintes conceitos:

Conceito A: -3, 0, -1, -2
Conceito B: -2, 1, 0, -1
Conceito C: 0, 3, 2, 1
Conceito D: 3, 1, 2, 3.

As correlações entre A, B e C são todas perfeitas, isto é, elas são iguais a 1,00. No entanto, o conceito A está muito mais próximo de B do que de C. Ademais, a correlação entre C e D é 0,00 e, no entanto, os dois conceitos estão bem próximos, como mostra a figura 6-6.

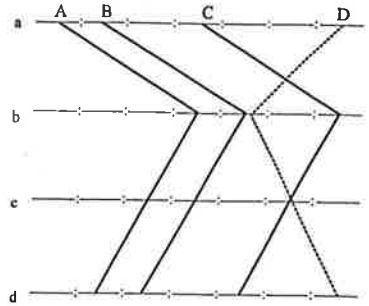


Fig. 6-6. Perfil de 5 conceitos (A, B, C, D) em 4 fatores (a, b, c, d)

Assim, é preciso deduzir uma fórmula que leve em conta as discrepâncias dos escores dos conceitos em cada fator e não somente a covariância entre estes escores. Tal tarefa é efetuada pela Fórmula Generalizada de Distância ("generalized distance formula"), que é a seguinte:

$$D_{AB} = \sqrt{\sum_j d^2_{AB}}$$

onde,

D_{AB} = distância semântica geral entre conceito A e B
d_{AB} = distância entre conceito A e B em cada fator.

Assim, por exemplo, as distâncias entre os conceitos pai (P), herói (H) e destino (D) da Tabela 6-3, serão as seguintes:

$$D_{PH} = \sqrt{(2,33-2,67)^2 + (1,33-2,67^2 + (0,67-1,67)^2} = 1,71$$

$$D_{PD} = \sqrt{(2,33--0,33)^2 + (1,33-1,33)^2 + (0,67-0,33)^2} = 2,68$$

$$D_{HD} = \sqrt{(2,67--0,33)^2 + (2,67-1,33)^2 + (1,67-0,33)^2} = 3,55$$

Ilustrando estas distâncias num gráfico, temos a figura 6-7.

Os dados da análise das distâncias mostram que há maior similaridade entre os conceitos de pai e herói e maior diferença entre herói e destino.

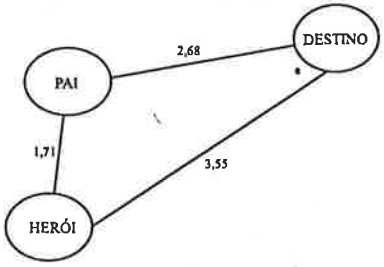


Fig. 6-7. Representação gráfica das distâncias entre três conceitos

3 - Usos do D

Se fizermos duas suposições, em princípio razoáveis, de que 1) as distâncias entre os pontos da escala de resposta que vai de -3 a +3 sejam iguais, isto é, que a escala é intervalar, e 2) que as respostas dos sujeitos nas diferentes escalas (itens) são independentes, então podemos considerar os escores do Diferencial Semântico como compondo uma medida métrica e todas as análises estatísticas permitidas com este tipo de escala numérica podem ser utilizadas. Desta forma, os escores D se prestam a uma série de análises, quais sejam,

a) Comparar a similaridade dos conceitos de um sujeito, como ilustra o exemplo acima dado da figura 6-6;
b) Comparar a similaridade de um mesmo conceito apresentado pelo mesmo sujeito em ocasiões diferentes, o que é útil, por exemplo, num contexto de psicoterapia ou treinamento;
c) comparar a similaridade de conceitos entre dois ou mais sujeitos;
d) comparar a similaridade de conceitos entre diferentes grupos de sujeitos. Neste caso, inclusive, como temos médias e desvios padrões, que resultam das distribuições de frequência dos escores dos sujeitos dentro de cada grupo, podemos até fazer teste de hipótese da diferença entre os conceitos através do teste "t", por exemplo.

4 - A Estrutura Conceitual

Em vez de comparar conceito com conceito, podemos ainda comparar todos os conceitos do Diferencial Semântico de uma só vez, calculando todas as diferenças n(n-1)/2 possíveis. Tal operação produzirá uma matriz que representa a estrutura semântica dos conceitos, a qual expressa as similaridades entre todos os conceitos e que pode ser ilustrada num gráfico de *n* dimensões, sendo *n* o número de conceitos envolvidos, como mostra a tabela 6-4 e a figura 6-7, trabalhando os dados da tabela 6-3.

Tabela 6-4. Distâncias D entre cinco conceitos

Conceitos	Pai	Herói	Destino	Guerra	Paz
Pai	-				
Herói	1,71	-			
Destino	2,68	3,55	-		
Guerra	5,51	5,83	2,99	-	
Paz	0,47	1,50	3,07	5,86	-

A ilustração que mostra a estrutura destes conceitos está na figura 6-8,

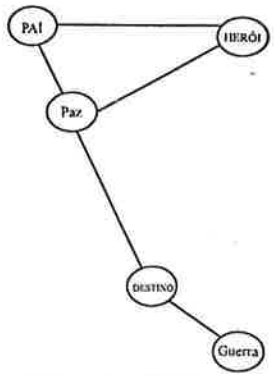


Fig. 6-8. Ilustração de uma estrutura bidimensional de 5 conceitos

A figura 6-8 mostra que os cinco conceitos apresentam uma estrutura composta de dois núcleos: um formado pelos conceitos pai, herói, paz e o outro por destino e guerra, sendo o conceito guerra o que mais destoa do grupo pai, herói e paz.

As escalas de tipo diferencial semântico têm-se mostrado bastante fidedignas, com índices de precisão teste-reteste variando entre 0,83 a 0,91 (Osgood e colaboradores, 1957), chegando até a 0,97 (Jenkins, Russell, & Suci, 1957). Osgood e colaboradores (1957) apresentam também altos índices de validade concorrente do Diferencial Semântico com as escalas de Thurstone (entre 0,74 e 0,82) e de Guttman (da ordem de 0,79).

Para o leitor brasileiro, há uma exposição prática da técnica de Osgood no livro de Alves Pereira (1986) pela Editora Ática de São Paulo.

Bibliografia

Alves Pereira, C.A. (1986). *O diferencial semântico. Uma técnica de medida nas ciências humanas e sociais.* São Paulo: Editora Ática.

Jenkins, J.J., Russell, W.A., & Suci, G.J. (1957). An atlas of semantic profiles for 360 words. In *Studies on the role of language in behavior.* Tech. Rep. No. 15. Minneapolis: University of Minnesota.

Morris, C.W. (1946). *Signs, language, and behavior.* New York: Prentice-Hall.

Osgood, C.E., Suci, G.J., & Tannenbaum, P.H. (1957). *The measurement of meaning.* Urbana, Ill.: University of Illinois Press.

Osgood, C.E. & Suci, G.J. (1952). A measure of relation determined by both mean difference and profile information. *Psychological Bulletin, 49,* 251-262.

Testes Referentes a Conteúdo: Medidas Educacionais

Luiz Pasquali
Amélia Regina Alves

Os testes referentes a conteúdo, como o nome deixa antecipar, dizem respeito a instrumentos centrados na análise de um conteúdo qualquer, caracterizando-se o teste por ser uma amostra representativa deste conteúdo, como ocorre tipicamente em provas educacionais. A legitimidade e o interesse de um tal instrumento surgem precisamente por ser uma amostra representativa do conteúdo definido, sendo aí também precisamente onde surgem os problemas e as dificuldades destes instrumentos. Para viabilizar um teste com validade de conteúdo é preciso que se façam as especificações do conteúdo que o teste pretende medir antes da construção dos itens, o que não é tarefa fácil. Estas especificações comportam a definição de três grandes temas: 1) definição do conteúdo, 2) explicitação dos processos psicológicos (os objetivos) a serem avaliados, e 3) determinação da proporção relativa de representação no teste de cada tópico do conteúdo.

A - Educação, Treinamento e Avaliação

A tecnologia de construção de testes referentes a conteúdo é lógica e tecnicamente bastante simples. Ela praticamente se reduz à questão da *tabela de especificação*, onde são enquadrados os objetivos da educação e os conteúdos do que avaliar. Os problemas graves nesta tecnologia se encontram precisamente nos procedimentos teóricos, onde se deve definir o que representa conteúdo e processos cognitivos, que tipicamente vêm definidos sob a égide de "objetivos" educacionais e que a tabela de especificação deve explicitar. A tarefa para tal definição se complica dada a diversidade de opiniões que existem sobre o que é educação e treinamento e o que se deve avaliar nos alunos. Os posicionamentos nesta área são os mais extremados e, inclusive, vêm influenciados por preocupações de caráter político e social. Assim, há os que defendem que a educação tem caráter seletivo: escolher os melhores para formação universitária e se tornarem os dirigentes da sociedade. Este enfoque vem desde os tempos antigos. Outros pensam que a educação deve ser um meio de estimular a qualidade de vida de todos os membros da sociedade. Num e noutro caso, discute-se o que a educação deve desenvolver nos sujeitos, argüindo-se, então, que processos cognitivos, habilidades, competências e outros valores humanos devem ser especialmente atendidos.

Por outro lado, o treinamento é consensuado por teóricos de Tecnologia Educacional (TED) como uma seqüência de experiências ou oportunidades destinadas a modificar o comportamento do indivíduo para atingir um objetivo declarado. Este conceito está estritamente relacionado à formação de habilidades, de sorte que o indivíduo seja considerado capaz para o desempenho de suas tarefas. Muitas organizações têm despendido consideráveis quantias de dinheiro retreinando seus empregados. Decorre daqui, igualmente, uma discussão prévia sobre quais as habilidades, competências ou conhecimentos a serem desenvolvidos.

Este capítulo não vai entrar nesta briga de caráter político, social ou cultural, mas cada pesquisador vai ter que tomar uma decisão quanto a este problema quando for construir um teste referente a conteúdo. Ele vai ter que definir o que ele entende por conteúdo, tanto curricular quanto de processos cognitivos, qualquer que seja o campo de atuação. Independentemente da decisão tomada, a tecnologia de construção de instrumentos referentes a conteúdo é sempre a mesma. Assim, a parte psicométrica destes instrumentos não é afetada pelo discurso sócio-político; este afeta sim e muito a qualidade dos instrumentos assim construídos em termos de pertinência. Um instrumento construído com base numa postura filosófica de educação como processo seletivo nunca irá satisfazer ao leitor que tem uma visão mais, digamos, humanista de educação. Isto significa que um instrumento pode ser tecnicamente (psicometricamente) válido, mas filosófica ou epistemologicamente inválido. Aliás, este problema é característico de qualquer procedimento teórico na construção de instrumentos psicossociais, pois as teorias nesta área são, além de abundantes, tipicamente insuficientes para explicar os fenômenos que estudam e são até contraditórias. As teorias sobre educação infelizmente não escapam a estes dilemas.

B - Planejamento Sistêmico da Instrução

Mesmo não querendo entrar numa discussão filosófica, não há como escapar de algumas considerações sobre o planejamento sistêmico da instrução e de seus componentes, sem cujo enquadramento não é possível este tipo de avaliação. A discussão será de caráter técnico, sem entrar na avaliação de valores educacionais.

B.1 - Conceituação

O planejamento sistêmico da instrução é a combinação organizada de elementos para o alcance de objetivos de ensino preestabelecidos. É formado por componentes que compreendem processos ou estratégias que interagem entre si para promover a mudança de performance. Sua finalidade básica é levar o indivíduo da condição de "não saber" para a condição de "saber". Suas principais características são:

- organização planejada e interdependente dos elementos;
- objetivos instrucionais claramente definidos;
- elementos organizados seqüencialmente de modo a facilitar a aprendizagem do aluno;
- "feed-back" ou sistema de avaliação contínua definida e freqüente.

B.2 - Componentes Básicos

Há três componentes básicos para um enfoque instrucional:

1. Objetivos instrucionais e natureza da aprendizagem: Tanto o planejador e o administrador da instrução quanto o aprendiz devem saber, claramente, o que deverá ser aprendido.
2. Estratégia instrucional: O "gerente de aprendizagem" determina a estratégia apropriada. Em alguns casos, o próprio aprendiz é quem determina como aprender; em outros, o professor e os materiais ajudam.
3. Avaliação: Esta deve estar medindo as habilidades especificadas pelos objetivos instrucionais. É um processo contínuo de coleta de dados. Dependendo da habilidade exigida dos alunos, ela envolve respostas cognitivas, afetivas ou motoras.

Neste capítulo serão detalhados mais os itens 1 e 3 do sistema instrucional integrado.

C - Objetivos Instrucionais e Natureza da Aprendizagem

A palavra e o conceito de objetivos, objetivos educacionais, objetivos pedagógicos, etc., são abusivamente utilizados no contexto da educação e do treinamento, onde mais se utilizam os testes referentes a conteúdo. Uma das conseqüências desta ocorrência consiste no fato de que *objetivos* constitui uma expressão recheada de ambigüidades ou conotações. Fala-se, inclusive, de objetivos a longo prazo, objetivos a curto prazo, objetivos gerais, objetivos específicos, objetivos instrucionais, objetivos intermediários, objetivos estratégicos, objetivos pedagógicos, objetivos comportamentais, etc. Na verdade, não há contradições entre todas estas acepções da palavra objetivos; entretanto, ela pode trazer dificuldades no momento em que se quer falar de avaliação educacional. Se esta consiste em verificar precisamente se foram ou não atingidos os objetivos, então fica inicialmente complicado saber do que é que estamos falando, quando dizemos que queremos medir os objetivos do ensino, por exemplo. Quando dizemos que o objetivo da educação é mudar o comportamento do educando ou que consiste em formar bons cidadãos, estamos falando de coisas diferentes. Mudar o comportamento é o objetivo estratégico, enquanto que formar bons cidadãos indica o conteúdo desta mudança de comportamento, os propósitos, os valores educacionais. Ambos são corretos, mas falam de coisas diferentes. Claro, se pretendo mudar o comportamento do educando, tenho que definir para onde ou em que sentido quero que ele mude, qual é o propósito de tal mudança desejada; e, no converso, se quero formar bons cidadãos, tenho que provocar mudanças nos educandos. Assim, quando for avaliar a mudança, tenho que avaliar se houve mudança e em que sentido ela ocorreu. Então, para poder viabilizar tal avaliação, tenho que conhecer os objetivos de conteúdo que serviram de orientação na intervenção que faço sobre os comportamentos do educando através da educação.

Para tal intento, uma taxionomia de objetivos se impõe. Mas destas existem mais do que se precisa. Entretanto, vamos tentar, pelo menos, esclarecer alguns conceitos nesta área para poder viabilizar a elaboração de testes de avaliação na mesma. Vamos abordar o problema de dois pontos de vista: um visa mostrar os tipos de distinções possíveis sob as quais os objetivos podem ser divididos ou

A figura 6-8 mostra que os cinco conceitos apresentam uma estrutura composta de dois núcleos: um formado pelos conceitos pai, herói, paz e o outro por destino e guerra, sendo o conceito guerra o que mais destoa do grupo pai, herói e paz.

As escalas de tipo diferencial semântico têm-se mostrado bastante fidedignas, com índices de precisão teste-reteste variando entre 0,83 a 0,91 (Osgood e colaboradores, 1957), chegando até a 0,97 (Jenkins, Russell, & Suci, 1957). Osgood e colaboradores (1957) apresentam também altos índices de validade concorrente do Diferencial Semântico com as escalas de Thurstone (entre 0,74 e 0,82) e de Guttman (da ordem de 0,79).

Para o leitor brasileiro, há uma exposição prática da técnica de Osgood no livro de Alves Pereira (1986) pela Editora Ática de São Paulo.

Bibliografia

Alves Pereira, C.A. (1986). *O diferencial semântico. Uma técnica de medida nas ciências humanas e sociais*. São Paulo: Editora Ática.

Jenkins, J.J., Russell, W.A., & Suci, G.J. (1957). An atlas of semantic profiles for 360 words. In *Studies on the role of language in behavior*. Tech. Rep. No. 15. Minneapolis: University of Minnesota.

Morris, C.W. (1946). *Signs, language, and behavior*. New York: Prentice-Hall.

Osgood, C.E., Suci, G.J., & Tannenbaum, P.H. (1957). *The measurement of meaning*. Urbana, Ill.: University of Illinois Press.

Osgood, C.E. & Suci, G.J. (1952). A measure of relation determined by both mean difference and profile information. *Psychological Bulletin, 49*, 251-262.

Testes Referentes a Conteúdo: Medidas Educacionais

Luiz Pasquali
Amélia Regina Alves

Os testes referentes a conteúdo, como o nome deixa antecipar, dizem respeito a instrumentos centrados na análise de um conteúdo qualquer, caracterizando-se o teste por ser uma amostra representativa deste conteúdo, como ocorre tipicamente em provas educacionais. A legitimidade e o interesse de um tal instrumento surgem precisamente por ser uma amostra representativa do conteúdo definido, sendo aí também precisamente onde surgem os problemas e as dificuldades destes instrumentos. Para viabilizar um teste com validade de conteúdo é preciso que se façam as especificações do conteúdo que o teste pretende medir antes da construção dos itens, o que não é tarefa fácil. Estas especificações comportam a definição de três grandes temas: 1) definição do conteúdo, 2) explicitação dos processos psicológicos (os objetivos) a serem avaliados, e 3) determinação da proporção relativa de representação no teste de cada tópico do conteúdo.

A - Educação, Treinamento e Avaliação

A tecnologia de construção de testes referentes a conteúdo é lógica e tecnicamente bastante simples. Ela praticamente se reduz à questão da *tabela de especificação*, onde são enquadrados os objetivos da educação e os conteúdos do que avaliar. Os problemas graves nesta tecnologia se encontram precisamente nos procedimentos teóricos, onde se deve definir o que representa conteúdo e processos cognitivos, que tipicamente vêm definidos sob a égide de "objetivos" educacionais e que a tabela de especificação deve explicitar. A tarefa para tal definição se complica dada a diversidade de opiniões que existem sobre o que é educação e treinamento e o que se deve avaliar nos alunos. Os posicionamentos nesta área são os mais extremados e, inclusive, vêm influenciados por preocupações de caráter político e social. Assim, há os que defendem que a educação tem caráter seletivo: escolher os melhores para formação universitária e se tornarem os dirigentes da sociedade. Este enfoque vem desde os tempos antigos. Outros pensam que a educação deve ser um meio de estimular a qualidade de vida de todos os membros da sociedade. Num e noutro caso, discute-se o que a educação deve desenvolver nos sujeitos, argüindo-se, então, que processos cognitivos, habilidades, competências e outros valores humanos devem ser especialmente atendidos.

Por outro lado, o treinamento é consensuado por teóricos de Tecnologia Educacional (TED) como uma seqüência de experiências ou oportunidades destinadas a modificar o comportamento do indivíduo para atingir um objetivo declarado. Este conceito está estritamente relacionado à formação de habilidades, de sorte que o indivíduo seja considerado capaz para o desempenho de suas tarefas. Muitas organizações têm despendido consideráveis quantias de dinheiro retreinando seus empregados. Decorre daqui, igualmente, uma discussão prévia sobre quais as habilidades, competências ou conhecimentos a serem desenvolvidos.

Este capítulo não vai entrar nesta briga de caráter político, social ou cultural, mas cada pesquisador vai ter que tomar uma decisão quanto a este problema quando for construir um teste referente a conteúdo. Ele vai ter que definir o que ele entende por conteúdo, tanto curricular quanto de processos cognitivos, qualquer que seja o campo de atuação. Independentemente da decisão tomada, a tecnologia de construção de instrumentos referentes a conteúdo é sempre a mesma. Assim, a parte psicométrica destes instrumentos não é afetada pelo discurso sócio-político; este afeta sim e muito a qualidade dos instrumentos assim construídos em termos de pertinência. Um instrumento construído com base numa postura filosófica de educação como processo seletivo nunca irá satisfazer ao leitor que tem uma visão mais, digamos, humanista de educação. Isto significa que um instrumento pode ser tecnicamente (psicometricamente) válido, mas filosófica ou epistemologicamente inválido. Aliás, este problema é característico de qualquer procedimento teórico na construção de instrumentos psicossociais, pois as teorias nesta área são, além de abundantes, tipicamente insuficientes para explicar os fenômenos que estudam e são até contraditórias. As teorias sobre educação infelizmente não escapam a estes dilemas.

B - Planejamento Sistêmico da Instrução

Mesmo não querendo entrar numa discussão filosófica, não há como escapar de algumas considerações sobre o planejamento sistêmico da instrução e de seus componentes, sem cujo enquadramento não é possível este tipo de avaliação. A discussão será de caráter técnico, sem entrar na avaliação de valores educacionais.

B.1 - Conceituação

O planejamento sistêmico da instrução é a combinação organizada de elementos para o alcance de objetivos de ensino preestabelecidos. É formado por componentes que compreendem processos ou estratégias que interagem entre si para promover a mudança de performance. Sua finalidade básica é levar o indivíduo da condição de "não saber" para a condição de "saber". Suas principais características são:

- organização planejada e interdependente dos elementos;
- objetivos instrucionais claramente definidos;
- elementos organizados seqüencialmente de modo a facilitar a aprendizagem do aluno;
- "feed-back" ou sistema de avaliação contínua definida e freqüente.

B.2 - Componentes Básicos

Há três componentes básicos para um enfoque instrucional:

1. Objetivos instrucionais e natureza da aprendizagem: Tanto o planejador e o administrador da instrução quanto o aprendiz devem saber, claramente, o que deverá ser aprendido.
2. Estratégia instrucional: O "gerente de aprendizagem" determina a estratégia apropriada. Em alguns casos, o próprio aprendiz é quem determina como aprender; em outros, o professor e os materiais ajudam.
3. Avaliação: Esta deve estar medindo as habilidades especificadas pelos objetivos instrucionais. É um processo contínuo de coleta de dados. Dependendo da habilidade exigida dos alunos, ela envolve respostas cognitivas, afetivas ou motoras.

Neste capítulo serão detalhados mais os itens 1 e 3 do sistema instrucional integrado.

C - Objetivos Instrucionais e Natureza da Aprendizagem

A palavra e o conceito de objetivos, objetivos educacionais, objetivos pedagógicos, etc., são abusivamente utilizados no contexto da educação e do treinamento, onde mais se utilizam os testes referentes a conteúdo. Uma das conseqüências desta ocorrência consiste no fato de que *objetivos* constitui uma expressão recheada de ambigüidades ou conotações. Fala-se, inclusive, de objetivos a longo prazo, objetivos a curto prazo, objetivos gerais, objetivos específicos, objetivos instrucionais, objetivos intermediários, objetivos estratégicos, objetivos pedagógicos, objetivos comportamentais, etc. Na verdade, não há contradições entre todas estas acepções da palavra objetivos; entretanto, ela pode trazer dificuldades no momento em que se quer falar de avaliação educacional. Se esta consiste em verificar precisamente se foram ou não atingidos os objetivos, então fica inicialmente complicado saber do que é que estamos falando, quando dizemos que queremos medir os objetivos do ensino, por exemplo. Quando dizemos que o objetivo da educação é mudar o comportamento do educando ou que consiste em formar bons cidadãos, estamos falando de coisas diferentes. Mudar o comportamento é o objetivo estratégico, enquanto que formar bons cidadãos indica o conteúdo desta mudança de comportamento, os propósitos, os valores educacionais. Ambos são corretos, mas falam de coisas diferentes. Claro, se pretendo mudar o comportamento do educando, tenho que definir para onde ou em que sentido quero que ele mude, qual é o propósito de tal mudança desejada; e, no converso, se quero formar bons cidadãos, tenho que provocar mudanças nos educandos. Assim, quando for avaliar a mudança, tenho que avaliar se houve mudança e em que sentido ela ocorreu. Então, para poder viabilizar tal avaliação, tenho que conhecer os objetivos de conteúdo que serviram de orientação na intervenção que faço sobre os comportamentos do educando através da educação.

Para tal intento, uma taxionomia de objetivos se impõe. Mas destas existem mais do que se precisa. Entretanto, vamos tentar, pelo menos, esclarecer alguns conceitos nesta área para poder viabilizar a elaboração de testes de avaliação na mesma. Vamos abordar o problema de dois pontos de vista: um visa mostrar os tipos de distinções possíveis sob as quais os objetivos podem ser divididos ou

dicotomizados; o outro visa dar uma breve visão de uma taxionomia clássica dos objetivos em termos do conteúdo (curricular e de processos).

Do ponto de vista dos tipos de objetivos possíveis, faz-se uma série de classificações, entre as quais é importante entender as seguintes dicotomias:

- *Mudança vs. Conteúdo*. Esta distinção realmente é entre objetivos (conteúdo, propósito da educação) e meta-objetivos (mudança de comportamento). Quando dizemos que a educação visa a mudança do comportamento do educando, estamos na realidade fazendo uma meta-análise do processo educativo, isto é, ele serve para produzir mudanças de comportamento. Agora, esta mudança pode ser orientada para as mais variadas direções, dependendo dos valores (o conteúdo, o propósito) que se quer modificar. Inclusive, neste sentido, torna-se objetivo legítimo e almejado a própria formação de peritos assassinos, como seria quiçá o interesse dos órgãos de segurança (veja o filme "Point of no Return" - A Assassina!). As taxionomias existentes sobre os objetivos se referem a esta última concepção de objetivos.
- *Objetivo Geral vs. Objetivos Específicos*. Esta distinção diz respeito à abrangência dos objetivos. Por exemplo, seria um objetivo geral da educação "formar bons cidadãos", enquanto um específico seria "ensinar a ler e escrever" ou "ensinar a ler um gráfico". Os objetivos gerais visam dar uma orientação genérica em direção à qual se quer que a mudança de comportamento dos educandos ocorra. Realmente são as orientações filosóficas sobre educação que ditam estes objetivos gerais, que tipicamente são tão genéricos que se pode entendê-los das maneiras mais diversas possíveis. Por exemplo, o que seria exatamente "formar bons cidadãos"? Cada leitor provavelmente vai entender tal objetivo de uma forma bem pessoal e diferente dos outros. Mas eles são importantes para, pelo menos, dar as balizas, as fronteiras amplas dentro das quais se situa o que educação deve ser. Para se poder avaliar tais objetivos é necessário torná-los específicos, isto é, operacionalizá-los de alguma forma em termos de comportamentos.
- *Objetivos Abstratos (Constitutivos) vs. Objetivos Comportamentais*. Os objetivos específicos podem ser expressos em termos abstratos ou em comportamentos. Trata-se realmente das definições dos objetivos, que podem ser constitutivas ou operacionais. Aquelas definem ou expressam as objetivos em termos de abstrações, ao passo que estas os expressam em comportamentos. Por exemplo, se meu objetivo for "aumentar no educando a compreensão do sistema numérico", este seria um objetivo abstrato, porque o que deve o educando fazer exatamente para mostrar que compreendeu? Ao passo que "somar números inteiros" é um objetivo comportamental, porque está claro quando se pede ao educando "vá e some estes números" (veja Mager, 1981).
- *Conteúdo vs. Processos*. A expressão "conteúdo" é bastante confusa no contexto da presente discussão, porque ela pode se referir a tudo que se discute em avaliação referente a conteúdo ou pode se referir apenas ao conteúdo programático de um curso ou treinamento, isto é, à matéria que se espera os educandos dominarem no final do citado curso. Tipicamente, o conteúdo nesta última acepção se divide em unidades e sub-unidades ou tópicos e sub-tópicos. Quando se contrapõe conteúdo a processos, quer se entender aquele como o programa de unidades e sub-unidades. Neste contexto, processos se referem aos processos

cognitivos e afetivos ou competências / habilidades que se quer desenvolver nos educandos no trabalho educativo. Se o conteúdo é dividido em unidades e sub-unidades, os processos são enquadrados dentro de taxionomias, das quais temos uma série delas clássicas, como as de Gagné (1963; 1965) e as de Bloom (1956; Bloom, Hastings, & Madaus, 1971; Krathwohl, Bloom, & Masia, 1964).

- *Conhecimento vs. Habilidades*. Conhecimento se refere à organização, armazenamento e recuperação de fatos, conceitos, princípios e procedimentos, enquanto habilidades se refere ao uso destes conteúdos em seqüências simples ou complexas, tais como desempenho musical ou artístico, raciocínio, solução de problemas (veja Haladyna, 1994). Snow e Lohman (1989) preferem fazer esta distinção em termos de conhecimento declarativo vs. conhecimento procedural, onde o primeiro representa o conceito acima definido de conhecimento e o segundo significando o uso do conhecimento declarativo para a consecução de um objetivo qualquer. Sob a influência da psicologia cognitiva, esta distinção vem recebendo atenção especial, sobretudo com referência aos processos psicológicos mais elevados (raciocínio, solução de problemas, conhecimento estratégico), mas o consenso nesta área ainda está longe de ser efetivo e praticamente útil para a avaliação educacional (veja Greeno, 1980; Crooks, 1988; Green, Halpin, & Halpin, 1990; Stiggins, Griswold, & Wikelund, 1989; Roid & Haladyna, 1982; Shavelson, Baxter, & Pine, 1992; Wiggins, 1989; veja também a discussão sobre "construct-driven tests" vs. "task-driven tests", Messick, 1989, 1994).

D - Taxionomia dos Objetivos Educacionais

Como dissemos, estamos falando aqui dos objetivos entendidos como os processos ou competências. Entre as várias taxionomias possíveis e existentes, vamos considerar brevemente as de Bloom (1956; Bloom, Hastings, & Madaus, 1971; Krathwohl, Bloom, & Masia, 1964), deixando para um Anexo, ao final deste capítulo, a exposição sobre outras taxionomias clássicas nesta área.

Para o seu autor, a estrutura de taxionomia que propõe parece favorecer a categorização dos mais diferentes objetivos educacionais. Ele define seu esquema taxionômico, partindo do princípio de que desempenhos simples, quando integrados com outros igualmente simples, se tornam mais complexos. Pode-se, então, dizer que as classificações representam estruturas onde comportamentos semelhantes se agrupam e formam a classe A (por exemplo, comportamentos de reconhecimento), outro conjunto de comportamentos a classe B (comportamentos de análise), etc.

A ordenação do comportamento simples para o mais complexo deve corresponder ao grau de dificuldade exigido. Assim, problemas que exigem os comportamentos da classe A tendem a ter maior freqüência de acerto e serão mais facilmente resolvidos do que os problemas que exigem os comportamentos da classe B, por exemplo. Problemas que requerem conhecimento de conceitos possuem maior freqüência de acerto do que os que requerem o conhecimento desses mesmos conceitos e mais a sua aplicação.

A aplicação de um processo taxionômico compreende, além de entendimento, certo nível de consciência. Razão pela qual, Bloom desenvolve, não somen-

te uma taxionomia de objetivos cognitivos, mas igualmente uma de objetivos afetivos. Para Bloom, a possibilidade de se constatar que o nível de consciência é um componente significativo na classificação de comportamentos demanda um conjunto de interesse de pesquisa para os psicólogos que atuam nesta área. Este ponto acaba de ser amplamente popularizado com o livro de Daniel Goleman (1996), "Inteligência Emocional".

A taxionomia de Bloom é a seguinte:

Domínio Cognitivo:

- *Conhecimento*: envolve as tarefas de *memorização* (relembrar, recordar material previamente aprendido). Se refere aos comportamentos de reprodução de
  - específicos (*elementos*): recordar termos e fatos específicos de informação
  - maneiras de lidar com específicos (*relações*): organizar, julgar, criticar a seqüência de fatos específicos (convenções, tendências e seqüências, classificações e categorias, critérios, metodologias)
  - universais e abstrações (*princípios*): conhecer esquemas e estruturas de como os fenômenos se organizam (princípios, teorias, generalizações, estruturas);
- *Compreensão*: apreender o que é comunicado, sem ver todas as conseqüências (entender o significado do material). Não se trata apenas de reproduzir, mas de entender o reproduzido; assim, o sujeito mostra capacidade para
  - *tradução*: transferir informação com precisão de uma forma para outra
  - *interpretação*: explicar ou sumariar uma comunicação
  - *extrapolação*: tirar conseqüências, implicações, etc. consentâneas com a comunicação original (estender o significado além dos dados);
- *Aplicação*: usar abstração em situações concretas (idéias, métodos e regras gerais, bem como princípios técnicos, idéias e teorias que devem ser lembradas e aplicadas). A este nível, o sujeito é capaz de fazer uso da informação em situações concretas da vida;
- *Análise*: Decompor uma comunicação em seus elementos ou partes constituintes, esclarecendo a hierarquia das idéias. Isto implica na demonstração de que o sujeito é capaz de identificar, numa comunicação, as partes constituintes, quais sejam
  - *elementos*: identificar os elementos em uma comunicação
  - *relações*: identificar conexões e inter-relações entre elementos
  - *princípios organizadores*: identificar a estrutura que mantém a comunicação uma unidade;
- *Síntese*: Integrar elementos e partes para formar um todo coerente e estruturado. Implica que o sujeito é capaz de
  - *produção de uma comunicação única*: desenvolver uma comunicação que comunique idéias, sentimentos, etc.
  - *produção de um plano ou um conjunto de operações*: desenvolver plano de trabalho
  - *derivação de um conjunto de relações abstratas*: desenvolver conjunto de relações abstratas que classifiquem ou expliquem dados, fenômenos, ou deduzir proposições e relações a partir de um conjunto de proposições básicas ou representações simbólicas;
- *Avaliação*: Julgar o valor de materiais e métodos para um dado propósito. O sujeito se mostra capaz de emitir julgamentos quantitativos e qualitativos sobre o quanto materiais e métodos satisfazem critérios. Segundo os critérios sejam dados ou lógicos, os julgamentos podem ser em termos de
  - *evidência interna*: avaliar em termos da adequação lógica, consistência, etc.
  - *critérios externos*: avaliar em termos de critérios selecionados ou recordados.

Domínio Afetivo:

- *Atenção*: sensibilizar o educando para certos fenômenos ou estímulos, isto é, que ele esteja disposto a receber ou atender a eles. Implica em
  - *consciência*: dar-se conta da presença do estímulo
  - *receptividade*: vontade de tolerar dado estímulo, não procurar evitar o estímulo
  - *atenção seletiva*: distinguir estímulos em figura e fundo; o educando já seleciona aspectos de estímulos ou estímulos que mais interessam a ele.
- *Responder*: o educando reage ao estímulo, procura-o e se sente feliz em trabalhar com ele. Isto se manifesta em comportamentos de
  - *aquiescer*: o educando reage ao estímulo, mas mais por obediência ou conforma-se com ele mais do que o aceita
  - *desejo de responder*: o educando mostra vontade de reagir ao estímulo, não por medo de punição, mas porque quer
  - *satisfação em responder*: além de querer reagir ao estímulo, o educando tem prazer em fazê-lo.
- *Valorizar*: o educando dá valor ao estímulo (tem atitude positiva), mostrando
  - *aceitação de um valor*: o educando considera o estímulo um valor com o qual se pode identificar
  - *preferência por um valor*: o educando escolhe valores e os procura ativamente
  - *compromisso*: o educando não somente aceita um valor mas se identifica com ele e tem convicção dele, procurando convencer outros sujeitos para se converterem à aceitação deste valor.
- *Organização*: na presença de mais de um valor, o educando é capaz de selecionar, hierarquizar os mesmos, isto é, ele é capaz de
  - *conceituar um valor*: o educando é capaz de ter uma visão abstrata dos valores, podendo assim ver suas relações
  - *organizar um sistema de valores*: capacidade de colocar junto uma série de valores, de preferência harmoniosamente ou, pelo menos, num equilíbrio aceitável.
- *Caracterização por um valor ou complexo de valores*: os valores se tornam organizados numa hierarquia e a vida do sujeito é regida por eles, formando a filosofia de vida e a sua concepção do mundo. Dois níveis são possíveis

- *conjunto geral*: os valores estão organizados num sistema interno dando consistência ao comportamento do indivíduo, definindo sua personalidade, caráter, atitudes
- *caracterização*: é o auge da organização dos valores numa hierarquia onde predomina um ou alguns valores e que pautam a vida do sujeito, dando a visão do mundo do sujeito.

Domínio Psicomotor

Mais ou menos na mesma época, Harrow (1972) desenvolveu um outro grupo de objetivos educacionais para complementar os citados acima. Estes objetivos incluem desde os reflexos motores até os movimentos de comunicação não-verbal. Embora tenha sido difícil classificar o domínio psicomotor, eis a taxionomia de Harrow como ilustração (útil sobretudo para disciplinas laboratoriais e para professores de educação física, dança, teatro, etc):

- *Movimentos Reflexos*: movimentos involuntários, que são subdivididos em reflexos segmentais, reflexos inter-segmentais, reflexos supra-segmentais.
- *Movimentos Básicos*: entram movimentos já mais complexos e movimentos que exigem certa habilidade. Seus sub-níveis são: movimentos locomotores, movimentos não-locomotores, movimentos manipulativos.
- *Habilidades Perceptivas*: se referem a todas as habilidades envolvidas no input de estímulos e que enviam as mensagens ao cérebro. Incluem os níveis de: habilidades quinestésicas, habilidades visuais, habilidades auditivas, habilidades táteis, habilidades coordenadas.
- *Habilidades Físicas*: habilidades desenvolvidas que vão constituir o self corpóreo do sujeito. Incluem resistência, força, flexibilidade, agilidade.
- *Movimentos Finos*: resultam da aprendizagem e contribuem para a eficiência na execução de tarefas motoras. Incluem habilidades simples, habilidades conjugadas, habilidades adaptativas complexas.
- *Comunicação Não-verbal*: os movimentos do corpo que constituem a linguagem do corpo, como expressões faciais, posturas, etc. Incluem movimentos expressivos e movimentos interpretativos.

E - Tabela de Especificação

E.1 - Conceituação e Estratégia

O elenco de objetivos que a taxionomia acima apresenta é um tanto assustador quando queremos intentar avaliá-los numa situação concreta (a avaliação educacional é precisamente isto que pretende fazer). Portanto, para se poder ter algum sucesso nesta empreitada de avaliar, alguma organização preliminar à construção dos itens da prova é absolutamente necessária, do contrário a avaliação fatalmente se tornará arbitrária e sem rumo. É o que se pretende fazer com o que Tyler (1950) chamou de "table of specification".

Esta tabela está baseada no fato de que cada objetivo instrucional é definido por dois componentes, a saber, o conteúdo instrucional (curricular) e os

comportamentos. O conteúdo instrucional constitui a matéria, o material, que é ensinado (conteúdo) e os comportamentos constituem aquilo que se quer que o educando faça com o material aprendido (processos cognitivos). Quem define o conteúdo, obviamente, é o especialista da matéria, ao passo que os processos são definidos por alguma taxionomia como a de Bloom que apresentamos acima.

Para viabilizar a construção de uma tabela de especificação, é preciso

1) estabelecer os objetivos gerais do currículo (tarefa do professor)
2) definir, para cada objetivo, o componente conteúdo e o componente processos
3) construir a tabela de especificação.

A tabela de especificação é uma matriz de duas dimensões, na qual uma das dimensões é encabeçada pelo conteúdo e a outra pelos comportamentos. Dentro da tabela estão definidos os itens que definem que conteúdo cobrir para avaliar que processo cognitivo, como no exemplo da tabela 7-1.

Tabela 7-1. Exemplo de tabela de especificação para um curso imaginário de estatística descritiva (no corpo está definido o número de itens)

Conteúdo	Processos (comportamentos)			Total
	Conceituar	Relacionar	Aplicar	
Freqüência	2	3	1	7
Tendência Central	3	1	5	9
Variabilidade	3	2	4	9
Total	8	6	10	25

A figura 7-1 dá um exemplo mais detalhado de uma tabela de especificação.

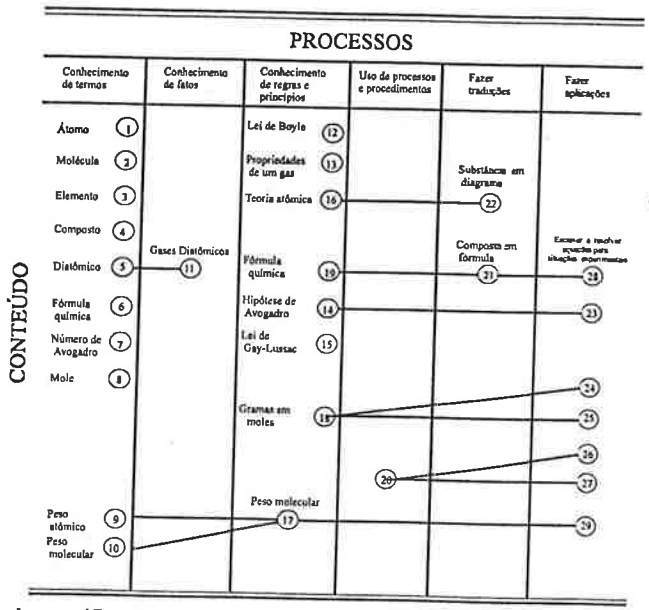


Fig. 7-1. Tabela de especificação para uma unidade de química (Bloom et al., 1971, p. 121)

E.2 - Cobertura dos Objetivos

Como os objetivos educacionais são inumeráveis, é preciso decidir que e quais objetivos introduzir na tabela de especificação para serem avaliados. Morris e Fitz-Gibbon (1978) oferecem umas regras práticas e úteis para formular e para selecionar objetivos educacionais para avaliação. Afirmam eles que constituem condições necessárias, para que eles possam ser avaliados, que os objetivos sejam

1. *Claramente formulados*. De fato, eles devem ser tão precisamente definidos que diferentes avaliadores possam construir tarefas (itens de avaliação) que sejam praticamente idênticas. Isto significa, na prática, que os objetivos a serem avaliados devem ser o mais específicos possível e preferencialmente operacionais (não abstratos) ou operacionalizados. Por exemplo, se o objetivo (abstrato) for "compreender a estatística descritiva", ele poderia ser melhor, isto é, mais claramente formulado, se for expresso como "saber calcular medidas de tendência central e de variabilidade".
2. *Atingíveis* num dado tempo e espaço. Longas listas de objetivos para um programa servem para fins de impressionar os órgãos financiadores, mas dificilmente será possível, num curso, implementá-los todos. Sendo este o caso, os objetivos que não foram cobertos no curso não fazem parte do curso e, com isso, não são atingíveis (atingidos) e, consequentemente, não podem ser avaliados. A formulação dos objetivos, segundo este requisito, deve ser feita em função da possibilidade de implementação do programa e não em função das imposições "oficiais" dos órgãos de fomento e secretarias de educação.
3. *Formulados ao nível de habilidade* que o programa quer desenvolver no aluno. Pela taxionomia de Bloom acima exposta, vimos que uma amplitude vasta existe de níveis de competências que um programa pode almejar, que vão desde a reprodução até a síntese e crítica. Então, a formulação dos objetivos deve especificar a quais destes níveis o ensino pretende levar. Por exemplo, se quiser, num curso de psicopatologia, que o aluno chegue a poder definir o que seja doença mental, não seria pertinente formular os objetivos em termos de como se cura a esquizofrenia. O converso também é verdadeiro, isto é, se quero obter do aluno a competência em curar a doença mental, não é suficiente expressar os objetivos em termos de compreender a doença mental. Um exemplo rústico:

   Objetivo: "Conhecer as partes constituintes da bicicleta".
   Teste: "Pegue esta bicicleta e ande 100 metros".

   O teste está avaliando um objetivo não presente e nem derivável do programa, pois não é necessário saber andar de bicicleta para saber quais são as partes componentes da mesma.
4. *Importantes*. Normalmente as listas de objetivos são intermináveis. Mesmo se todos estes intermináveis objetivos forem importantes, acontece que dificilmente eles todos podem ser avaliados num dado momento. O que fazer? Duas soluções possíveis:
   4.1 *Amostrar os objetivos* a avaliar, isto é, selecionar uma amostra dos objetivos para serem avaliados. Morris e Fitz-Gibbon (1978) propõem quatro métodos para se conseguir tal amostra, como aparece na tabela 7-2, os quais são: priorizar objetivos por pontos, priorizar objetivos por hierarquização, amostrar objetivos randomicamente, amostrar objetivos por blocos.

Tabela 7-2. Técnicas de amostragem de objetivos educacionais

Método	Especificação	Vantagem	Desvantagem	Recomendações
Priorizar por pontos	Grupo de 15 avaliadores pontuam todos os objetivos numa escala de 5 pontos. A média da pontuação define as prioridades	- produz prioridades exatas segundo grupos de interesse - envolve outros avaliadores e cria assim credibilidade - assegura relevância da avaliação em termos de valores do cliente	- focaliza avaliação sobre número reduzido de objetivos - depende da cooperação de avaliadores - demanda tempo para avaliação e dos avaliadores	Boa técnica se há tempo suficiente e cooperação de avaliadores. Especialmente útil se avaliação deve ser sensível ao programa e aos valores e desejos do grupo cliente
Priorizar por hierarquia	Agrupar objetivos por áreas; organizá-los do simples ao complexo. Os mais complexos recebem maior prioridade	- pode ser feito pelo avaliador sozinho - focaliza objetivos complexos e desfocaliza objetivos simples demais	- exige tempo, sobretudo se os objetivos são numerosos - nem sempre interessa avaliar só objetivos terminais ou difíceis - a matéria pode não permitir hierarquia	Viável quando o avaliador deve organizar os objetivos ele mesmo e a matéria não oferece uma hierarquia lógica
Amostra randômica	Selecionar randomicamente os objetivos a partir de sua totalidade	- pode ser feito pelo avaliador sozinho - método simples e rápido - trata todos os objetivos do programa como importantes	- perigo omitir objetivos importantes e periclitar a credibilidade da avaliação	Recomendável quando os objetivos são de igual importância
Amostra por blocos	Objetivos são alocados em partes do teste e cada parte é dada a grupos diferentes de alunos, estes tipicamente selecionados randomicamente	- pode ser feito pelo avaliador sozinho - muitos objetivos podem ser avaliados, embora com alunos diferentes - evita cópia de respostas, pois alunos tomam testes diferentes	- procedimento complexo - dados não se prestam a certas análises estatísticas - não prático com crianças para quem se deve ler as instruções - não permite comparar os alunos	Único método possível para avaliar muitos objetivos. É complexo, daí usá-lo quando os outros métodos não funcionam

Fonte: Morris & Fitz-Gibbon, 1978, ps. 28-29

4.2 *Generalizar os objetivos*. Em vez de avaliar diretamente os objetivos específicos formulados no programa, produzir objetivos mais gerais (para avaliar), os quais se pode supor que o programa desenvolveu nos alunos. Seriam como uma espécie de objetivos de segunda ordem, os quais congregariam sob si vários dos objetivos primários explicitamente formulados no programa, como ilustra o esquema da figura 7-2.

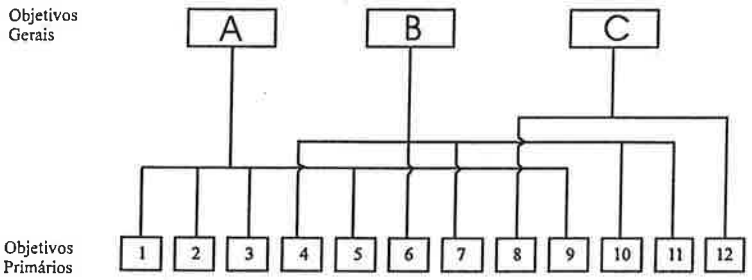


Fig. 7-2. Objetivos gerais (transferência) e objetivos primários em avaliação

Para elaborar estes objetivos mais gerais, faça os seguintes questionamentos em cima dos objetivos do programa:

- que habilidades o aluno deve ter desenvolvido após o programa - habilidades para as quais o programa supostamente preparou o aluno?
- que habilidades deveriam se desenvolver nos alunos para as quais o programa poderia indiretamente contribuir?
- que outras habilidades a escola espera que os alunos desenvolvam ao serem submetidos a este programa?

F - Técnicas de Avaliação dos Objetivos de Domínio Cognitivo

F.1- Preliminares: Para viabilizar a construção de um teste de conteúdo, é preciso efetivar alguns passos, tais como,

1. Dividir o conteúdo em unidades
2. Definir cada unidade em seus componentes de
   - conteúdo e
   - comportamentos
3. Definir em cada unidade os elementos
   - essenciais (importantes)
   - secundários
4. Incluir na tabela de especificação todos os elementos essenciais
5. A especificação da unidade deve incluir itens que avaliem todos os níveis comportamentais ou aqueles de interesse do avaliador (os vários processos cognitivos da taxionomia)
6. Definir o(s) formato(s) dos itens, tais como, múltipla escolha, complementação de sentenças, etc. O mesmo teste pode ter vários formatos de itens.

F.2 - Tipos de Tarefas para Avaliar Aprendizagem

Os itens ou tarefas para avaliar a aprendizagem podem ser expressos em vários formatos, que serão expostos a seguir. De um modo geral, a taxionomia do formato das tarefas pode ser reduzida a dois tipos, a saber, escolher uma resposta ou construir uma resposta. No primeiro tipo, o sujeito precisa apenas selecionar a resposta correta dentre um conjunto de respostas possíveis oferecidas para uma questão; no outro tipo, o sujeito deve redigir a resposta correta ele mesmo. Cada um destes dois tipos possui um elenco de sub-tipos como mostra o esquema abaixo (Haladyna, 1994):

Resposta Selecionada	Resposta construída
Escolha múltipla	Complementação de sentença
Emparelhamento	Ensaio curto
Verdadeiro - falso	Ensaio extenso

1. *Múltipla escolha*: consiste em escolher uma (correta) entre várias alternativas apresentadas de resposta. Itens deste tipo constam de três partes: uma base (uma pergunta ou frase incompleta), uma resposta certa, e vários distratores (respostas erradas).

Exemplo:

Traço latente em Psicologia é valorizado por
a. behavioristas
b. cognitivistas
c. estatísticos
d. humanistas
e. ambos *b* e *d*. *

Recomendações:

- cada item deve cobrir apenas um problema e bem formulado na base, de sorte que basta ler a base para saber a resposta correta, sem ver as alternativas
- fazer questões e alternativas curtas
- assegurar que haja apenas uma resposta certa ou uma melhor resposta
- as alternativas erradas têm que ser plausíveis e atrativas ao aluno menos informado
- variar a localização da resposta correta na ordenação das alternativas
- enunciar a base em forma positiva: evitar questões negativas (geralmente são confusas)
- incluir suficientes alternativas (cerca de 5) para minimizar acertos por mero acaso (chute)
- para tornar a questão difícil, faça as alternativas o mais similares possível
- evite usar, como alternativas, "todas elas" ou "nenhuma delas"
- faça que cada alternativa tenha a mesma estrutura e tamanho.

Esta técnica é muito versátil e permite cobrir grande quantidade de material, além de ser simples a correção (ela é objetiva). Entretanto, para evitar o chute é preciso que o número de alternativas seja grande e isto introduz uma nova dificuldade, a saber, que as alternativas falsas devem apresentar aparência de verdadeiras, isto é, não devem ser estúpidas ou obviamente erradas, inclusive para os menos avisados. Dados estes e os outros requisitos mencionados nas recomendações, a

elaboração de questões de múltipla escolha demanda tempo, muito embora esse tempo possa ser compensado no momento da correção. Haladyna (1994) oferece 43 regras práticas para elaborar itens de resposta selecionada, especialmente de múltipla escolha. Veja estas regras na tabela 7-3.

Tabela 7-3. As 43 regras de redação de itens

Procedimentos Gerais
1 - Use o formato de resposta correta ou da melhor resposta
2 - Evite itens de múltipla resposta complexa (tipo K)
3 - Formate o item verticalmente, não horizontalmente
4 - Destine tempo para editar e outras revisões dos itens
5 - Use boa gramática, pontuação e ortografia
6 - Minimize o tempo de leitura ao escrever cada item
7 - Evite itens capciosos, que enganam ou iludem o respondente a dar resposta errada
Conteúdo dos itens
8 - Tome como base para o item um objetivo educacional ou instrucional
9 - Focalize apenas um único problema
10 - Mantenha o vocabulário consistente com o nível de compreensão do respondente
11 - Mantenha os itens independentes, evitando que um dê dicas para o outro
12 - Use exemplos para desenvolver seus itens
13 - Evite conhecimentos ou assuntos super-especializados
14 - Evite frasear o item verbatim a partir do manual do curso
15 - Evite itens baseados em opiniões
16 - Use itens de múltipla-escolha para avaliar processos mais elevados de pensar
17 - Teste material importante, significativo; evita itens triviais
Construção da base
18 - Escreva a base em formato de questão e não de complementação
19 - Ao escrever itens de complementação, não deixe os brancos nem no começo nem no meio da base
20 - Assegure-se que a base seja clara e diga exatamente o que está querendo perguntar
21 - Evite enfeitar a base com verbosidade excessiva
22 - Enuncie a base positivamente; frases negativas são confusas
23 - Inclua na base a idéia central e a maior parte do fraseado
Construção das alternativas
24 - Use tantas alternativas quanto possível
25 - Coloque as alternativas em ordem lógica ou numérica
26 - Mantenha as alternativas independentes; elas não devem se sobrepor
27 - Mantenha todas as alternativas de um item homogêneas em conteúdo
28 - Mantenha o comprimento das alternativas consistente
29 - Evite ou use com parcimônia a frase "todas as acima"
30 - Evite ou use com parcimônia a frase "nenhuma das acima"
31 - Evite a frase "não sei"
32 - Formule as alternativas positivamente, não negativamente
33 - Evite dicas para os sabidos, tais como, opções absurdas, associações, etc.
34 - Evite dar dicas através da construção gramatical incorreta
35 - Evite determinantes específicos, tais como, "nunca', "sempre"

Construção da alternativa correta
36 - Coloque a alternativa correta de modo que apareça proporcionalmente em todas as posições de alternativas
37 - Assegure-se de que haja uma e apenas uma alternativa correta
Construção das alternativas falsas
38 - Use alternativas plausíveis; evite alternativas ilógicas
39 - Use, nas alternativas falsas, erros comuns dos alunos
40 - Use termos técnicos nas alternativas erradas
41 - Use frases familiares mas incorretas nas alternativas falsas
42 - Use afirmações verdadeiras, mas que não respondem ao item
43 - Evite uso de humor nas alternativas erradas

Anotações à:

- *regra 2*: exemplo de tal item seria
  O que afeta a fidedignidade do teste?
  1 - Comprimento do teste
  2 - Homogeneidade da amostra de sujeitos
  3 - Comprimento do item
  A - 1 e 2 *
  B - 2 e 3
  C - 1 e 3
  D - 1, 2, e 3
- *regra 3*: exemplo de tal situação seria
  Qual seria um exemplo de satélite?

Certo
A - Terra
B - Lua
C - Sol
D - Ambos A e B

Pior			
A - Terra	B - Lua	C - Sol	D - Ambos A e B

2. *Verdadeiro - Falso*: Consiste em dizer se a frase é verdadeira (correta - V) ou errada (falsa -F). Exemplos:
   a. Rogers era behaviorista
   b. Freud definiu o conceito de reforço
   c. Precisão é um parâmetro dos testes psicológicos.

Recomendações:

- o enunciado deve cobrir um problema apenas
- as questões têm que ser claramente verdadeiras ou falsas, sem qualificações
- frases curtas
- ter mais ou menos o mesmo número de questões verdadeiras e falsas num teste
- misturar as questões verdadeiras e falsas

- as questões verdadeiras e falsas devem ter mais ou menos a mesma extensão
- evitar enunciados negativos.

Esta técnica permite cobrir grande quantidade de material, são fáceis de corrigir, mas difíceis de construir já que, exigindo uma resposta absoluta de verdadeiro ou falso, seu enunciado deve ser inquestionavelmente verdadeiro ou falso. Contudo, elas apelam quase exclusivamente para a memória bruta e permitem muito chute (dá para acertar por acaso em 50% das vezes), por isso se exige um teste com muitos itens. Dado este índice alto de chute, algumas vezes se sugere um desconto por cada resposta errada (penalização por chute), sendo o sujeito convidado a omitir (deixar em branco) as questões cuja resposta correta ele desconhece.

3. *Emparelhamento (correspondência)*: são apresentadas duas colunas de informações, uma de descrições ou premissas (A) e outra de opções ou respostas (B); as informações de B devem ser pareadas com as de A. Exemplo:

Escreva na frente da teoria psicológica expressa em A a letra do seu correspondente fundador indicado em B.

A	B
1 — Behaviorismo	a. Skinner
2 — Humanismo	b. Rogers
3 — Cognitivismo	c. Dollard e Miller
4 — Frustração e Agressão	d. Katz e Kahn
5 — Dissonância Cognitiva	e. Sternberg
	f. Freud
	g. Spearman
	h. Festinger.

Recomendações:

- cada item de correspondência deve cobrir material homogêneo, para que todas as opções apareçam razoáveis
- as listas a parear devem ser de frases curtas
- todas as opções devem ser plausíveis
- a lista de descrições (A) deve ter as frases mais longas; as opções (B) frases curtas
- enumere as descrições e identifique as opções com letras
- inclua mais opções que descrições
- permita que uma opção possa ser utilizada mais de uma vez
- na instrução, dê a regra de emparelhamento.

Esta técnica evita o chute e as questões são de fácil construção, mas permite avaliar quase exclusivamente memorização (reconhecimento).

4 - *Complementação de sentenças*: consiste em preencher lacunas de frases incompletas ou produzir uma frase. Exemplos:

- O nome do criador da teoria da relatividade é _____;
- O behaviorismo foi fundado por _____;
- O psicólogo e estatístico _____ foi um dos pioneiros da Psicometria;
- Dê a definição de inteligência segundo Spearman;
- Dê os autores das seguintes teorias psicológicas

a. behaviorismo _____
b. psicanálise _____
c. humanismo _____

- 4 + 12 - 6 = ?.

Recomendações:

- preferir questões cuja resposta exige uma única palavra
- a resposta exigida deve ser claramente correta
- palavra a completar deve ser importante, não uma irrelevante
- deixar em branco palavras de modo a que a frase não perca sentido, como por exemplo: "Brasília é a _____ do Brasil e fica _____ no _____"; seria difícil alguém descobrir a frase toda como "Brasília é a capital do Brasil e fica localizada no Centroeste"
- é melhor colocar a palavra eliminada lá pelo fim da sentença
- se a resposta é em números, dizer as unidades a usar.

Esta técnica facilita a construção e correção das questões, mas pode ocasionar ambigüidade na correção, sobretudo se frases inteiras forem produzidas pela educando.

5 - *Ensaio* (essay): O ensaio é um teste em que é dada a mais ampla liberdade de resposta ao aluno. O professor coloca a pergunta e o estudante decide como atacar o problema, que informação utilizar, como organizar a matéria e que importância dar a cada aspecto do tema.

Costuma-se distinguir dois tipos de teste de ensaio, a saber:

1) *ensaio curto*: ensaio com resposta limitada. Às vezes, a pergunta por si só pode pôr limites de resposta, como, por exemplo, "enumere as 6 características de X"; outras vezes, o enunciado do ensaio põe os limites, como, por exemplo, "descreva e justifique, em duas laudas, o enunciado "a fidedignidade do teste é condição necessária mas não suficiente da validade do teste";
2) *ensaio longo*: ensaio livre, no qual nenhum limite de espaço, número de páginas, etc. é colocado e onde o sujeito tem total liberdade de responder da maneira que quiser. É ótimo teste para medir os mais altos objetivos do domínio cognitivo, mas de extrema dificuldade de correção.

O ensaio constitui uma técnica altamente recomendada para avaliar as habilidades mais complexas do domínio cognitivo, por exemplo, quando se pretende apreciar o estilo e a criatividade do sujeito. Ela, contudo, apresenta sérios problemas (Gronlund, 1974):

1) não é capaz de medir adequadamente o aproveitamento. O ensaio não tem como se constituir numa amostra representativa do conteúdo de uma matéria, uma vez que ele pode cobrir apenas algumas áreas. Estas ele avalia profundamente, mas o restante da matéria não é avaliado e, portanto, não se está avaliando o aproveitamento total de um curso, por exemplo. Também por isso, o ensaio pode ser bastante injusto para os estudantes, pois vai receber um escore alto quem ou estudou/aprendeu bem toda a matéria ou aquele que estudou/aprendeu bem apenas

aquele tópico da matéria que por acaso entrou como tema do ensaio. Um outro estudante que aprendeu bem 90% da matéria, apenas não justamente aquela que o ensaio cobriu, recebe um escore de fracasso no curso;

2) depende intrinsecamente da capacidade de escrever. Assim, quem tem dificuldades em escrever está quase fatalmente fadado a um resultado inferior num ensaio, ainda que conheça perfeitamente a matéria, enquanto um aluno hábil em escrever, mas sem maiores conhecimentos do conteúdo, pode se sair bem, blefando, com o estilo de escrever, a falta de conhecimento do conteúdo;
3) é de difícil correção, pois esta é amplamente subjetiva, dependendo dos gostos, valores e pendências do corretor, além de demandar enormes quantidades de tempo.

Diante das qualidades e das dificuldades do teste de ensaio, é importante seguir algumas regras de como elaborar tais instrumentos:

1) Utilize ensaio somente para medir resultados complexos de aprendizagem. Na taxionomia dos objetivos, o ensaio é ideal para avaliar as habilidades de síntese e de avaliação; é bem menos interessante para avaliar o conhecimento, a aplicação e a análise;
2) Enuncie o problema a ser elaborado de uma maneira bem clara, definida e o mais restrita possível. Assim, seria estranho para o ensaio um enunciado do tipo "fale sobre validade dos testes". Seria bem mais pertinente enunciar "defina, descreva, explique [ou outro verbo desta natureza] o conceito de validade dos testes psicológicos no contexto da psicometria clássica";
3) De preferência, não dê aos estudantes a opção de poder escolher entre diferentes temas para o mesmo ensaio, pois isso dificulta comparar o resultado, já que estamos normalmente procurando avaliar a aprendizagem de um grupo;
4) Dê tempo suficiente, mas não ilimitado para responder à pergunta de ensaio. É bom informar periodicamente aos estudantes do andamento do tempo durante a prova.

Há igualmente algumas regras úteis para a correção de ensaios, a saber:

1) Avalie as respostas dadas no ensaio em termos da aprendizagem do tema que se está medindo. Dar pontos para ortografia e caligrafia é legítimo, mas o peso principal do escore deve cair sobre o tratamento dado ao tema em termos dos conhecimentos do mesmo pelo aluno;
2) Sobretudo em ensaios de responsa restrita é normalmente possível elaborar um crivo de correção contendo as informações ou dados que a resposta deve mencionar, inclusive já dando pesos diferenciados a estas informações;
3) Em ensaios de resposta livre, uma técnica bastante comum consiste em fazer uma leitura da prova e qualificá-la numa escala de cinco níveis de qualidade. Isto pode ser feito sobre a qualidade total da prova ou sobre cada critério previamente estabelecido, como, por exemplo, 1) integridade do plano (se o texto tem começo, meio e fim), 2) clareza e precisão na exposição de cada passo, 3) propriedade da justificativa de cada argumento;
4) É melhor trabalhar pergunta por pergunta do ensaio do que aluno por aluno, o que propicia um critério mais uniforme de avaliação das respostas. Comparar as respostas dos alunos na mesma pergunta evita melhor o erro de halo do que quando

se avalia cada resposta do mesmo sujeito em função das respostas dadas por ele a todas as questões;

5) As análises das respostas dos diferentes alunos devem ser feitas anonimamente, sem identificação do respondente;
6) Deve haver mais de um avaliador para uma mesma prova ou cuidar do acordo dos avaliadores previamente.

6 - *Os Portfólios*: Os portfólios vêm assumindo um papel cada vez mais relevante no contexto da avaliação na escola; eles pretendem ser uma alternativa às avaliações educacionais tradicionais e insistem que a avaliação do rendimento na escola não pode ser efetuada unicamente com provas escolares, mas que nesta avaliação se deve levar em conta toda a história acadêmica do aluno e todos os produtos que ele desenvolveu no mesmo período. Assim, para se ter um aferimento justo das potencialidades e do desempenho, o aluno deve ser avaliado, levando em conta todas as suas produções, sejam elas de ordem acadêmica, "hobbies", atividades extracurriculares, etc.; enfim, os portfólios contêm toda a história acadêmica do aluno, isto é, praticamente toda a sua biografia, o seu "curriculum vitae" acadêmico.

A filosofia que fundamenta esta visão parece justa e meritória, a saber, a pessoa deve ser avaliada e valorizada em termos de toda a sua vida e não exclusivamente em termos de alguns pontos isolados e esporádicos da vida, e exclusivamente em termos de desempenho cognitivo, como são as provas escolares. A dificuldade com os portfólios, contudo, se situa na sua implementação prática e válida. Na verdade, como avaliar adequadamente as atividades extracurriculares, por exemplo; e que pesos dar às diversas atividades do aluno? Têm elas todas o mesmo peso, igual, digamos, à aferição da aprendizagem formal em sala de aula? A tecnologia dos portfólios suscita ainda mais dúvidas que soluções, pois até o presente ela não se encontra razoavelmente delineada e operacionalizada. Nem por isso e, talvez por causa disso, ela deve ser perseguida, pesquisada e adequadamente operacionalizada, pois ela não se reduz a avaliar o aluno exclusivamente em termos de habilidades cognitivas, quase sempre puramente abstratas, mas em termos de todo o potencial e competências do mesmo, nas mais variadas situações escolares e, até, da vida.

F.3 - Técnicas para Avaliar Conhecimento

Itens (tarefas) que avaliam conhecimento fazem apelo à memória e ao reconhecer (aprendizagem bruta e verbalização - Henry, 1946). Implica em evocar (produzir: evocação) ou reconhecer (reconhecimento) um dado apresentado.

1 - Regras

- cobrir conteúdos de fato ensinados em sala de aula
- utilizar terminologia usada nos procedimentos do próprio ensino

2 - Técnicas

Assim, os itens não podem apelar para generalizações ou vocabulário desconhecido pelo educando, pois trata-se de recordar e não de analisar ou aplicar

conhecimentos. Exemplo: Todos os exemplos dados em E.2 são de conhecimento.

Ações típicas que definem este nível de habilidades são: definir, descrever, listar, identificar, etiquetar, parear, nomear, recordar, recitar, selecionar, declarar, esboçar, distinguir entre, exemplificar (dar um exemplo).

F.4 - Técnicas Para Avaliar Compreensão

Na hierarquia da taxionomia dos objetivos, compreensão se encontra a um nível acima de conhecimento e engloba as seguintes três habilidades:

- *Tradução*: uma mensagem ou conceito conhecido é expresso com nova simbologia ou em outras palavras, diferentes daquelas em que ele foi ensinado. Por exemplo, interpretar com palavras um gráfico ensinado.
  → Trabalha com *elementos* da mensagem.
- *Interpretação*: consiste em relacionar as partes de um todo; diferenciar o essencial (importante) do secundário numa mensagem.
  → Trabalha com *relações* entre os elementos da mensagem.
- *Extrapolação*: consiste em ver ou tirar conseqüências possíveis a partir da mensagem, seja para o futuro ou para outro tema.
  → Trabalha com *conseqüências* que podem ser derivadas, mas vão além da própria mensagem.

A característica fundamental dos itens que medem compreensão consiste em que eles não sejam expressos nos exatos materiais com que a unidade foi ensinada, embora deva manter características similares em termos de linguagem, simbologia, nível de complexidade e dificuldade e de conteúdo.

Ações típicas destas habilidades são: converter, defender, distinguir, discriminar, calcular, explicar, expandir, generalizar, sumarizar, inferir, prever, parafrasear.

Exemplos: Para

Tradução: o exemplo implica tradução de uma idéia ou de uma forma verbal para uma forma gráfica. O exemplo poderia ser igualmente o contrário, isto é, dado o gráfico e o aluno escolher entre uma série de alternativas verbais oferecidas.

1 - Foi dado um teste de conhecimento bastante fácil a um grupo de estudantes. Qual dos três gráficos da Figura 7-3 explicaria melhor os resultados, onde *f* significa freqüência de estudantes e *e* o escore dos mesmos?

2 - Qual dos seguintes termos tem um significado mais parecido com o termo taxionomia?
   A. Classificação *
   B. Elaboração
   C. Avaliação
   D. Tradução
   E. Síntese.

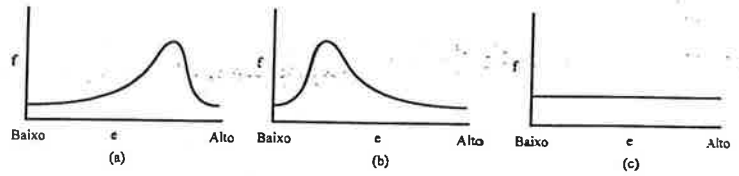


Fig. 7-3

Interpretação:

1 - As características Psicológicas dos seres humanos dependem de dois fatores, a saber, por um lado, da hereditariedade, que indica quanto elas dependem da herança genética e da maturação, e, por outro, do meio ambiente, físico, cultural e outros em que o sujeito foi criado.

Com base no gráfico da Figura 7-4, responda qual é a característica psicológica que
   a. mais depende do meio ambiente _____
   b. menos depende do meio ambiente _____
   c. depende igualmente do meio ambiente e da hereditariedade _____.

2 - O enunciado "a fidedignidade do teste é condição necessária mas não suficiente de sua validade" significa que
   A. Um teste fidedigno possui certo grau de validade
   B. Um teste válido possui certo grau de fidedignidade*
   C. Um teste fidedigno pode ser completamente inválido
   D. Um teste válido pode ser completamente impreciso
   E. Fidedignidade e validade são conceitos equivalentes.

Extrapolação:

1 - Com base na Figura 7-4, responda as seguintes questões: Qual das características
   a. se desenvolveria mais com uma boa educação _____
   b. menos aproveitaria de um treinamento ou exercício _____
   c. mais aproveitaria da existência simultânea de uma boa herança genética e de um meio ambiente favorável _____.

2 - O que é mais provável que ocorra com a fidedignidade de um teste de múltipla escolha, quando o número de alternativas de respostas mudar de 3 para 5?
   A. Aumentará *
   B. Diminuirá
   C. Permanece igual.

Exemplo Complexo: Sobre os mesmos dados se pode fazer questões que avaliam todos os três níveis de compreensão.

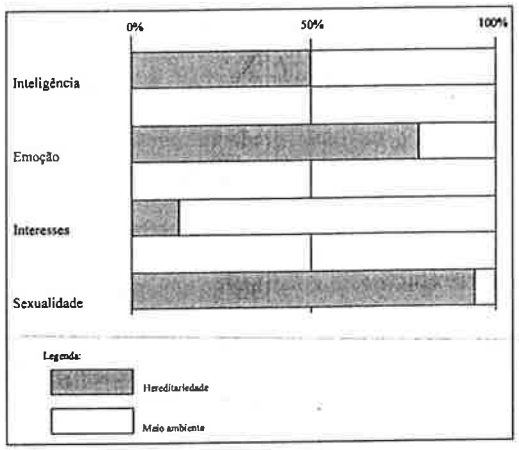


Fig. 7-4. Características psicológicas e sua dependência da hereditariedade e do meio ambiente

Com base nos dados da Figura 7-5, responda as questões que seguem:

1 - Qual dos grupos de sujeitos apresenta maiores escores em valores religiosos? (Tradução)
2 - Os estudantes femininos mostram maiores escores em que valores? (Tradução)
3 - Em que valor(es) os três grupos de sujeitos mais se assemelham? (Interpretação)
4 - Quais são os dois valores cujos escores mais se distanciam no caso dos estudantes masculinos? (Interpretação)
5 - Com base nos escores de valores, sujeito de que grupo seria uma escolha mais acertada para fundar uma empresa de negócios? (Extrapolação)
6 - Se você fosse iniciar um movimento em prol dos favelados, para que grupo de sujeitos seria mais compensador pedir apoio? (Extrapolação).

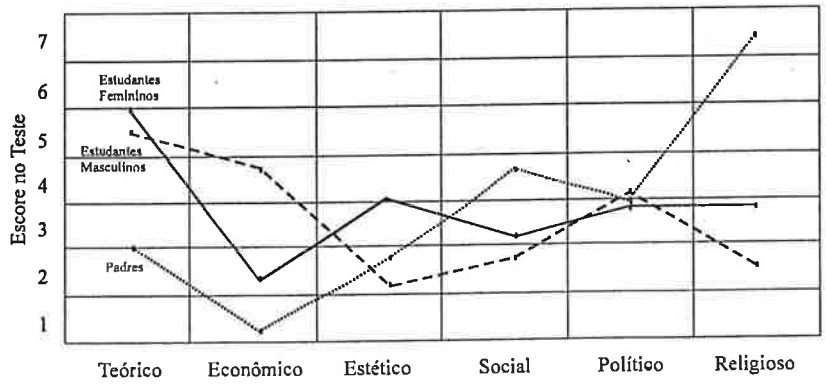


Fig. 7-5. Num estudo de valores, um grupo de estudantes universitários masculinos e femininos e de padres receberam os escores expressos na figura, numa escala de intensidade de 7 pontos (dados fictícios)

F.5 - Técnicas para Avaliar Aplicação

1 - Conceituações

A expressão "aplicar princípios e fazer generalizações" constitui um dos objetivos mais apreciados pelos educadores. Neste caso, o aluno não somente demonstra que compreendeu o significado da informação, mas que é igualmente capaz de aplicá-lo a situações concretas e novas. Esta habilidade vem expressa das mais variadas formas possíveis, mas de um modo geral ela implica "no uso de abstrações em situações particulares e concretas. As abstrações podem aparecer na forma de idéias gerais, regras de procedimentos e métodos generalizados. Elas podem igualmente se apresentar como princípios técnicos, idéias e teorias que precisam ser recordadas e aplicadas" (Bloom, 1956, p. 205). Os seguintes representam alguns exemplos de expressão de objetivos de aplicação:

- A habilidade de apresentar idéias [oralmente ou por escrito] de acordo com os princípios da gramática (Tyler, 1954);
- A habilidade de aplicar generalizações e conclusões da ciência social para problemas sociais atuais (Bloom, 1956);
- Habilidade para aplicar fatos e princípios científicos (French, 1957).

Qualquer que seja o conteúdo, os objetivos de aplicação vêm sempre expressos mais ou menos nos seguintes termos: *Habilidade para aplicar princípios e generalizações a novos problemas e situações*. Há aqui três construtos que podem produzir ambigüidades e, por isso, devem ser melhor esclarecidos e operacionalizados. São eles

- habilidade para aplicar
- princípios e generalizações
- problemas e situações novos.

Bloom e colaboradores (1971) esclarecem estes conceitos da seguinte forma:

1) *Problemas e situações novos*: o problema é novo se ele mantém similaridade com o que foi ensinado mas possui elementos diferentes, de sorte que não é suficiente ao estudante apelar para a memória para poder obter a resposta ao mesmo. Para ser novo, o problema não pode ser o mesmo que foi resolvido em sala de aula, apenas expresso com outras palavras; ele, embora sendo similar ao problema resolvido em aula, deve conter elementos diferentes, de sorte que o estudante, para poder resolvê-lo deve apelar para uma das seguintes alternativas: a) a expressão do problema deve ser modificada de algum modo para o problema poder ser resolvido (o aluno deve ver o que é essencial e secundário na expressão; deve perceber o que é dado e o que falta na expressão; talvez o problema deva ser expresso diferentemente); b) o problema deve ser colocado na forma de algum paradigma ou modelo, antes de procurar a solução; c) o problema exige que o estudante procure, em sua memória, princípios e generalizações que podem ser aplicados no caso.
2) *Princípios e generalizações*: De um modo geral, em qualquer assunto ou matéria existem algumas idéias básicas com as quais se pode dar um sumário de praticamente tudo que a ciência aprendeu sobre o mesmo. Estas idéias básicas de-

vem ser o objetivo principal de ensino, porque elas permitem ter uma visão o mais completa possível do campo de estudo e com elas se pode resolver praticamente todos os problemas ligados ao respectivo campo de saber. Algumas destas idéias básicas já estão bem estabelecidas e definidas, são os *princípios* ou *leis* que a ciência descobriu e que afirmam algo de fundamental sobre a realidade; por exemplo, a lei do reforço em Psicologia (um ato se repete se for reforçado), a lei da inércia em Física (um corpo se mantém na sua posição ou movimento até que uma força aja sobre ele) etc. *Generalizações* são afirmações gerais ou inferências amplas que se fazem dentro de um assunto qualquer e que permitem serem aplicadas em muitas situações similares, mas não em todas. Talvez a diferença maior entre princípios e generalizações está em que aqueles valem sempre, ao passo que estas últimas possuem muitas exceções. Por exemplo, "frustração produz ansiedade" é uma generalização mais do que um princípio, pois muitos outros fatores, com ou sem a frustração, também produzem ansiedade; mas, de um modo geral, é verdade que com a frustração ocorre ansiedade. Estas generalizações podem se basear em dados de pesquisa científica (como esta da frustração e ansiedade) ou podem ser resultados da pura observação, tais como, "fumar produz câncer" ou "crianças que recebem amor e atenção se desenvolvem melhor".

De qualquer forma, tais princípios e generalizações permitem ao estudante ter uma visão mais completa e ampla do campo de estudo e com eles pode resolver grande parte dos problemas que se apresentam neste campo. Eles próprios não são eventos singulares, mas permitem entender uma série de eventos sem mesmo ter que aprender um por um.

3) *Habilidade para aplicar*: isto implica que o estudante sabe utilizar os princípios e generalizações apropriados em situações novas, isto é, diferentes das que foram utilizadas em sala de aula para aplicar os mesmos princípios. Esta habilidade se manifesta em comportamentos do estudante, que podem ser sumarizados como segue (Bloom et al., 1971, p. 165): O estudante
   A) determina que princípios ou generalizações são apropriados e pertinentes ao presente problema;
   B) pode redefinir o problema de modo a tornar claro que princípios ou generalizações são necessários para sua solução;
   C) consegue especificar os limites dentro dos quais um dado princípio ou generalização é verdadeiro ou relevante;
   D) consegue ver as exceções de uma dada generalização e as razões delas;
   E) consegue explicar fenômenos novos em termos de princípios ou generalizações conhecidos;
   F) consegue prever o que acontecerá numa nova situação pelo uso de princípios e generalizações apropriados;
   G) sabe determinar ou justificar o curso de uma ação ou decisão numa nova situação pelo uso de princípios e generalizações apropriados;
   H) consegue expor a lógica que ele utiliza para usar este ou aquele princípio ou generalização para uma situação-problema nova.

2 - Técnicas para Testar a Aplicação (Exemplos)

Ações típicas dessas habilidades são: mudar, computar, demonstrar, desenvolver, empregar, modificar, operar, organizar, preparar, produzir, relatar, resolver, transferir, usar.

Um prova para testar a aplicação de princípios e generalizações pressupõe que

a) a situação-problema (a tarefa, o item, a questão) seja nova ou diferente daquela utilizada no ensino. Quanto mais diferente, mais difícil se torna a questão;
b) o problema apresentado deve ser solucionável, ao menos em parte, através do uso de princípios ou generalizações apropriadas;
c) o problema deve exigir uma ou mais das habilidades listadas acima (A - H).

2.1 - Testando habilidades A e B: não precisa resolver inteiramente o problema, basta definir os princípios ou generalizações apropriados para sua solução. Exemplo:

Marque o princípio que melhor explica cada afirmação de fatos. Diante dos fatos escreva a letra do princípio que se aplica.

Princípios explicativos:

A. Força é igual a massa vezes aceleração
B. O momentum de um corpo tende a permanecer constante
C. O momento ou efeito de torção de uma força é proporcional à distância do eixo de rotação
D. Existe fricção entre corpos em contato e se movimentando um com respeito ao outro
E. A soma das energias cinética e potencial num sistema isolado é uma constante.

Fatos:

1) tesouras usadas para cortar chapas de metal possuem longas manetas.
2) A força exercida num freio pelo pé do motorista é bem menor do que a exercida sobre os discos do freio.
3) Um foguete pode se impulsionar no vácuo.
4) Se uma mó que está rodando rapidamente se romper, os fragmentos voam para fora em linha reta.
5) fazer um automóvel aerodinâmico reduz o montante de força necessária para manter uma velocidade de 80 quilômetros por hora.

2.2 - Testando as habilidades C e D: reconhecer os limites de aplicação e as exceções na aplicação de um princípio ou generalização. Exemplo:

Nos seguintes itens, você encontra um fato seguido por uma conclusão. Avalie a conclusão, utilizando a escala de três pontos (A a C).

Escala:

A. O fato constitui uma boa evidência para *dar suporte* à conclusão.
B. O fato constitui uma boa evidência para *desaprovar* a conclusão.
C. Nem A nem B.

1) *Fato*: As tribos nativas da Austrália possuem uma organização social complexa e uma tecnologia muito simples.
   Conclusão: A complexidade de uma cultura não material não depende de um grande desenvolvimento tecnológico.
2) *Fato*: O número de acionistas na maioria das grandes corporações cresceu consideravelmente nos últimos 30 anos.
   Conclusão: O controle das corporações se tornou mais democrático nos últimos 30 anos.
3) *Fato*: A máquina tem estimulado uma crescente divisão de trabalho com ênfase em operações cada vez mais minuciosamente detalhadas.
   Conclusão: O trabalhador individual na sociedade moderna tipicamente não possui "insight" na totalidade do mecanismo social.

2.3 - Testando a habilidade E: Explicar fenômenos novos em termos de princípios e generalizações conhecidos. A explicação assume freqüentemente a forma de "A ocorre por causa de Y", onde Y é o princípio ou generalização invocado. Exemplo:

1) Se se levanta freqüentemente a tampa de uma chaleira na qual se está esquentando a água, esta demora mais para ferver porque
   A. Ferver ocorre a temperaturas mais altas se a pressão é aumentada
   B. O vapor que escapa leva consigo calor do líquido
   C. Deixando o vapor escapar diminui o volume do líquido
   D. A temperatura de um vapor é proporcional ao seu volume a temperatura constante
   E. Permitindo mais ar entrar resulta em aumento de pressão sobre o líquido.
2) Os degraus que levam para uma piscina parecem torcidos quando eles entram na água. Qual das seguintes é a melhor explicação para tal fenômeno?
   A. Deflação da luz pela superfície da água
   B. Dispersão da luz ao entrar a água
   C. Refração da luz devida à diferença na velocidade da luz no ar e na água
   D. A luz não caminha em linhas retas na água
   E. Partículas suspensas na água.

2.4 - Testando a habilidade F: Prever resultados ocorridos ou futuros fazendo uso de princípios ou generalizações conhecidos. Exemplo:

1) Suponha que um elevador está descendo numa aceleração de gravidade constante "g". Se um passageiro jogar uma bola de borracha para cima, qual será o movimento da bola com relação ao elevador? A bola
   A. Fica parada no ponto em que o passageiro a largou
   B. Sobe para cima do elevador e fica lá
   C. Não sobe nada, mas cai no chão
   D. Sobe, retorna e se move para o chão numa velocidade constante
   E. Sobe, retorna e se move para o chão numa velocidade crescente.
2) Marque a resposta (a, b, c) e dê a razão do porquê (A a E) para o seguinte problema: Um corpo de ar com temperatura de 40° possui uma umidade relativa de

40%. Se a temperatura do ar for aumentada para 60° sem acrescimento de água, a umidade relativa será
   a) aumentada
   b) diminuída
   c) inalterada

Porque,

   A. A capacidade do ar em manter água diminui com o aumento da temperatura
   B. A capacidade do ar em manter água é independente da temperatura do ar
   C. A capacidade do ar em manter água aumenta com o aumento da temperatura
   D. A taxa de evaporação aumentará
   E. A razão do peso do vapor de água para o peso do ar permanece a mesma.

2.5 - Testando a habilidade G: Tomar decisões com base em princípios e generalizações conhecidos. É especialmente útil em ciências sociais. Exemplo:

Muitas pessoas prevêem um período de forte inflação seguindo uma guerra. Se você fosse um consultor do Ministro das Finanças, diante da situação acima descrita, qual seria seu conselho com respeito às políticas abaixo descritas? Use a escala A e B para responder.

Escala:

   A. Aconselharia, se fosse para diminuir a tendência descrita acima
   B. Não aconselharia, se fosse para diminuir a tendência descrita acima.

Políticas:

1) Aumentar a taxa de desconto
2) Aumentar as facilidades para que mais firmas pudessem tomar dinheiro emprestado em termos justos
3) Fazer com que as pessoas descontassem seus papéis do tesouro imediatamente
4) Impor poupança forçada
5) Expandir programas públicos de trabalho
6) Diminuir drasticamente as taxas de juros.

2.6 - Testando a habilidade H: Além de usar princípios e generalizações, o estudante deve dizer porque os utilizou. Exemplo:

O reservatório de água de uma grande cidade é obtido de um grande lago, e o esgoto é despejado num rio que sai do lago. Houve um tempo em que este rio desembocava no lago, mas durante o período glacial a direção do rio foi invertida. Ocasionalmente, durante chuvas pesadas nas nascentes, águas do rio fluem para dentro do lago. O que se deve fazer para salvaguardar, de uma forma eficaz e econômica, a saúde das pessoas que vivem na cidade?

Tarefa:

Escolha a conclusão que você crê ser a mais consistente com os fatos dados acima e a mais racional segundo seus conhecimentos. Justifique sua conclusão, utilizando uma ou mais das razões expressas abaixo.

Conclusões:

A. Durante a estação da chuva, o montante de substâncias químicas usadas para purificar a água deve ser aumentado
B. Deve-se providenciar um sistema permanente de tratamento do esgoto antes de ser lançado no rio
C. Durante a estação da chuva, a água deve ser captada de um ponto do lago distante da origem do rio.

Razões:

1) Dado que as bactérias não podem sobreviver em carne salgada, podemos dizer que elas não podem sobreviver em água com cloro.
2) Muitas bactérias de esgoto não são prejudiciais ao homem.
3) Por cloro na água é um dos métodos menos caros para eliminar bactérias de um reservatório de água.
4) Um pessoa esclarecida saberia que a melhor maneira para matar bactérias é o uso de cloro.
5) Um sistema de tratamento de esgoto é mais barato que o uso de cloro.
6) Bacteriologistas afirmam que a melhor maneira de controlar bactérias é através do uso de cloro.
7) Na medida em que aumenta o número de microorganismos na água, a quantidade de cloro necessário para matar os organismos deve ser aumentado.
8) Um sistema de tratamento de esgoto é o único meio conhecido através do qual se pode garantir que água seja absolutamente sadia.
9) Aumentando o montante de cloro na água, a saúde das pessoas desta cidade estará protegida.
10) Bactérias perigosas na água são mortas quando uma pequena porção de cloro é colocado nela.
11) Quando as bactérias entram em contato com água com cloro, elas saem da água clorada para poder sobreviver.
12) Esgoto não tratado contém grande quantidade de bactérias, muitas das quais produzem doenças no homem.
13) Em muitas cidades é costumeiro se usar cloro para controlar bactérias perigosas nos reservatórios de água.
14) Esgoto depositado no lago tende a permanecer no local perto de onde foi lançado.

F.6 - Técnicas para Avaliar Análise

1 - Conceituação de Análise

Analisar consiste em decompor uma comunicação em seus elementos ou partes constituintes. Isto implica na identificação de

- Elementos (partes): idéias, valores, suposições, pontos de vista...
- Relações entre as partes, como hipótese → evidência, pressupostos → argumento, causa → efeito, seqüência lógica de relações...
- Forma como os elementos estão organizados (Princípios organizacionais): organização das partes, arranjo sistemático e estrutura.

Exemplos:

Análise de elementos:
- reconhecer termos básicos e suas relações; diferença entre afirmações subjetivas e objetivas;
- distinguir a conclusão da evidência que a suporta;
- reconhecer pressupostos não afirmados (suposições implícitas);
- identificar motivos e discriminar entre mecanismos de comportamento com referência a indivíduos e grupos;
- reconhecer técnicas defensivas e não defensivas usadas para influenciar pensamento e comportamento: propaganda, boatos, estereótipos, apelos emocionais, etc.

Análise de relações:
- reconhecer termos básicos e suas relações; o problema ou questão; o argumento do autor; evidência de suporte; o objetivo e pressuposições do autor;
- ver a consistência entre hipóteses e informação e suposições dadas;
- reconhecer que fatos ou suposições são essenciais para uma tese ou argumento em apoio da tese;
- identificar suposições não expressas necessárias para a linha de argumentação;
- reconhecer que detalhes são relevantes para a validação de um julgamento;
- reconhecer as relações causais e os detalhes importantes e secundários num relato histórico;

Análise de princípios organizacionais:
- inferir o objetivo ou ponto de vista do autor ou linhas de pensamento e sentimentos exibidos na sua obra;
- reconhecer o tom, temperamento e objetivo do autor;
- analisar, especialmente numa obra de arte, a relação dos materiais e meios de produção com os elementos e com a organização;
- reconhecer o ponto de vista ou viés de um escritor num conto histórico;
- reconhecer diferentes métodos de pesquisa científica, tais como classificação, pesquisa correlacional, pesquisa de causa e efeito, etc.

2 - Habilidades para analisar

Elementos:

A. O estudante pode classificar palavras, frases e afirmações num documento usando critérios analíticos oferecidos;

B. O estudante consegue inferir qualidades ou características particulares que não estão diretamente mencionadas, mas a partir de dicas disponíveis no documento;

Relações:

C. O estudante consegue inferir, a partir de critérios e relações do material num documento, que qualidades, suposições ou condições estão implícitas ou necessárias;

Princípios Organizacionais:

D. O estudante consegue utilizar critérios (tais como, relevância, causação, seqüência) para discernir um padrão, uma ordem ou arranjo do material num documento;

E. O estudante consegue reconhecer os princípios ou padrões de organização sobre os quais um documento inteiro está baseado;

F. O estudante consegue inferir a estrutura particular, o propósito e o ponto de vista nos quais o documento está baseado.

3 - Testando as Habilidades de Análise

Ações típicas dessas habilidades são: separar, deduzir, diagramar, diferenciar, distinguir, ilustrar, inferir, dar um esboço, mostrar, relatar, subdividir, retirar.

3.1 - Testando as Habilidades A: Reconhecer e classificar os elementos da comunicação (ver a função, o objetivo e o uso destes elementos). Exemplo:

1) Responda as questões abaixo usando a escala de A a E

Escala:

A - É *fato* e tem sido verificado *verdadeiro* por experimentos ou observação
B - É *fato* e tem sido verificado *falso* por experimentos ou observação
C - É *parte* de um teoria aceita
D - É *contraditório* com uma teoria aceita
E - É verdadeiro por simples *definição* de uma palavra ou uso de palavras

Questões:

1) A água congela a 0º centígrados.
2) O interior de um átomo é em sua maioria espaço vazio.
3) A pressão exercida por um gás se deve ao peso de moléculas.
4) O ferro enferruja em combinação com oxigênio.
5) Um volume igual de gases a uma mesma temperatura e pressão tem peso igual.
6) A resistência de dois condutores em série é maior do que sua resistência em paralelo.

2) Responda as questões abaixo usando a escala de A a D (pode usar as apostilas e anotações)

Escala:

A - A afirmação é uma hipótese que o pesquisador procura investigar em seu estudo
B - A afirmação é uma suposição na qual o estudo se baseia, mas não uma hipótese sobre a qual o estudo foi concebido
C - A afirmação é uma descoberta do estudo, mas não foi uma hipótese sobre a qual o estudo foi concebido
D - Nenhuma das acima.

Questões:

1) As atitudes sociais e econômicas de um indivíduo estão intimamente relacionadas à identificação de sua classe e ambos estão associadas à sua função no sistema econômico.
2) Algumas pessoas da classe média são mais radicais, nos termos definidos pelo autor, do que pessoas da classe trabalhadora.
3) A maioria dos brasileiros se consideram membros da classe dos trabalhadores.
4) A classe social de um indivíduo pode ser determinada com razoável segurança pelo montante e constância do seu salário.
5) A reação a questionários por parte de membros de classes sociais diferentes são suficientemente comparáveis de sorte que respostas diferentes podem ser tomadas como representando opiniões diferentes.
6) Uma classe social é caracterizada por atitudes e crenças comuns.
7) É possível declarar questões envolvendo atitude sócio-econômica liberal ou conservadora em termos que possuem o mesmo significado para todos os brasileiros.
8) Resultados obtidos de um amostra de 1.100 sujeitos são validamente generalizáveis para toda a população brasileira.

3.2 - Testando a Habilidade B: Descobrir elementos, qualidades ou características não explicitamente declaradas na comunicação, usando dicas nela constantes. Exemplo:

1) Na discussão de Leibniz sobre a "qualidade de movimento", seu primeiro postulado estabelece
   A. A definição do termo "força" adquirido por um corpo que cai de uma altura A.
   B. A relação entre corpos que caem a corpos projetados para cima contra a gravidade.
   C. Que o momentum adquirido por um corpo que cai de uma altura *h* é suficiente para levá-lo de volta à altura *h*.
   D. A equivalência do peso e da força motriz.
2) Seu segundo postulado estabelece
   A. A definição do termo "força".
   B. A relação entre altura de queda e velocidade adquirida.

C. A relação entre altura de queda e peso do corpo.
D. O caso especial que surge em considerando as máquinas.

3) Discutindo sobre a separação de partículas, Lavoisier *não* afirma ou assume que
A. Qualquer corpo expandido por calor pode ser contraído por esfriamento.
B. Existe um nível atingível de temperatura abaixo do ponto no qual os corpos permanecem constantes em tamanho apesar de resfriamento ulterior.
C. O tamanho das partículas individuais fica inalterado pelo calor.
D. Existe um ponto na escala de temperatura abaixo do qual as marcas se tornam sem sentido.

3.3 - Testando a Habilidade C: descobrir elementos que afetam toda a comunicação. Exemplo:

1) Responda as seguintes questões, usando a escala A e B

Escala:
A - a conclusão segue logicamente;
B - a conclusão não segue logicamente.

Questões:
(1) Declaração: se X existe, então Y existe. X existe.
Conclusão: Y existe.
(2) Declaração: se X existe, então Y existe. X não existe.
Conclusão: Y não existe.

2) Escolha a melhor resposta para as seguintes questões:
(1) Declaração: Nenhum amante do sofismo respeita a verdade. Todos os céticos amam o sofismo.
Conclusão:
A. Todos os céticos respeitam a verdade.
B. Alguns céticos respeitam a verdade.
C. Nenhum que respeita a verdade são não-céticos.
D. Alguns céticos não respeitam a verdade.
E. Nenhum delas.
(2) Declaração: É verdade que, se existe competição perfeita, o custo de produção inevitavelmente se iguala ao preço de venda. Contudo, competição perfeita nunca existiu, não existe e nem existirá.
Comentários:
A. A conclusão segue logicamente.
B. É impossível ter certeza de que a competição nunca existiu e nunca existirá.
C. Não está afirmado que a competição perfeita seja a única condição que permite o custo de produção ser igual ao de venda.
D. Um argumento que começa com "se" nunca pode levar a uma conclusão certa.
E. O argumento parece plausível, mas contem uma importante falácia.

3.4 - Testando Habilidades D, E, F: perceber a estrutura, o padrão, o ponto de vista de uma comunicação. Exemplo (avaliando uma obra de arte presente diante do aluno):

Habilidade D: A sensação de massa no trabalho [uma escultura] resulta de
A. O material.
B. A cor.
C. A forma.
D. O tratamento de superfícies.

Habilidade E:
1) Um dos principais movimentos na escultura é criado por
A. Movimento paralelo de frente e fundo dos planos da figura.
B. Um único mover cilíndrico de cima para a base.
C. Uma linha contínua circulando a figura.
D. Uma linha vertical estendendo de cima para a base.
2) O meio afeta a obra principalmente
A. limitando a natureza do desenho.
B. por sua cor e textura.
C. por sua estrutura e dureza.
D. pela dificuldade de esculpir.

Habilidade F: Esta obra é melhor descrita como
A. Representação fiel do objeto natural.
B. Representação seletiva do objeto natural.
C. Abstração baseada em princípios geométricos.
D. Quase não-objetiva.

F.7 - Técnicas para Avaliar Síntese

1 - Conceituação

Síntese consiste em ajuntar partes ou elementos para constituir um todo organizado. Faz parte ou, pelo menos, exige criatividade ("pensamento divergente").

2 - Habilidades de Síntese

A - Produzir uma comunicação única: o autor expressa (oral ou por escrito) idéias, sentimentos, relações ou experiências a outrem numa comunicação integrada (poema, pintura, composição musical, argumento matemático...).
B - Produzir um plano ou conjunto de operações: o autor produz um delineamento que poderá ser desenvolvido por ele mesmo ou outrem.
C - Produzir um conjunto de relações abstratas: o autor produz um modelo teórico para explicar dados ou fenômenos ou deduz proposições (hipóteses) a partir de suposições básicas ou representações simbólicas.

Ações típicas destas habilidades são: categorizar, compilar, compor, criar, delinear, formular, rescrever, sumarizar, planejar.

Exemplos:

Comunicação única:
Habilidade em escrever, usando boa organização de idéias e frases
Habilidade em escrever uma estória, verso, etc. com criatividade
Expressa suas idéias em discurso, por escrito ou outra forma artística, com clareza e correção
Habilidade em participar eficazmente em discussões de grupo sobre problemas sociais ... coordenando diferentes pontos de vista, sugerindo soluções e orientando-as para os objetivos do grupo
Habilidade em construir ... materiais gráficos

Produção de delineamento:
Habilidade em propor modos para testar hipóteses
Habilidade em planejar uma unidade de ensino para fins específicos
Habilidade em desenhar um edifício de acordo com especificações dadas.

Derivação de relações abstratas:
Desenvolvimento de uma hipótese explicativa a partir dos dados disponíveis
Habilidade em formular hipóteses baseadas numa análise dos fatores envolvidos e em modificá-las diante de novos fatores e considerações
Habilidade em fazer descobertas matemáticas e generalizações
Habilidade em perceber possíveis maneiras em que experiências podem ser organizadas para formar uma estrutura conceitual.

3 - Testando Síntese A:

Produzir uma composição, um "essay", uma redação sobre um tema, tendo em conta a platéia à qual ela se destina, o efeito que se quer obter sobre a mesma e a técnica a usar, isto é, oral, por escrito, por pintura, etc.

4 - Testando Síntese B:

Produzir um plano (delineamento) para resolver um problema. Não se trata de resolver o problema, mas definir procedimentos, etapas, dados necessários, critérios a seguir, para que o problema possa ser resolvido por alguém. Exemplo:

Elaborar um plano para a construção de um teste de temperamento, levando em conta

A. Teoria(s) do temperamento
B. Técnicas de construção de itens
C. Validação do teste
D. Normatização do teste
E. Tipos de análises estatísticas sugeridas em cada etapa

5 - Testando Síntese C:

Produzir um conjunto de hipóteses consistentes ou explicações para compreensão de um dado fenômeno (desenvolver um modelo ou teoria explicativa). Exemplo:

> Imagine que você está no futuro e vai estudar a cultura do Brasil no ano 3.000. Você descobre que a maioria dos postos de honra e poder são ocupados por mulheres. Ao perguntar às pessoas, você descobre que os atributos ideais de um indivíduo são a inteligência, a bondade e o respeito pelas obras criativas. Descobre, ainda, que as mulheres sobressaem aos homens em todos estes atributos. Faça uma redação descrevendo que outras mudanças sociais significativas acompanhariam as mudanças acima descritas.

Como o processo de síntese faz exigências sobre a capacidade do estudante para produzir e organizar idéias novas e originais, estando intimamente ligado à habilidade de escrever (Goldshalk, Swineford, & Coffman, 1966), este processo é tipicamente avaliado através de testes tipo ensaio.

F.8 - Técnicas para Avaliar Avaliação

1 - Conceituação:

Avaliar significa emitir julgamentos de valor sobre idéias, obras, material, etc., usando critérios pessoais ou dados pelo professor. Os critérios podem ser

- internos: consistência lógica dos julgamentos na comunicação
- externos: coerência dos julgamentos com critérios estabelecidos a priori pelo próprio aluno ou por especialista da área.

Exemplos:

Julgamentos por critérios internos: Habilidade em
- reconhecer a exatidão, completude e relevância de dados
- distinguir entre inferências, generalizações, argumentos e implicações válidas e inválidas
- reconhecer falhas, contradições e redundância num dado conjunto de postulados e detectar falácias em argumentos matemáticos.

Julgamentos por critérios externos: Habilidade em
- aplicar padrões estéticos previamente estabelecidos na escolha e uso de objetos comuns do meio ambiente
- reconhecer qualidade artística em obras de arte e música modernas
- uma proposição sobre fenômenos naturais
- comparar teorias, generalizações e fatos sobre uma dada cultura
- detectar e pesar julgamentos e valores envolvidos na escolha de uma linha de ação
- avaliar crenças tradicionais, instituições e padrões de conduta com relação à função do Estado.

2 - Habilidades de Avaliação:

A - Julgar um documento ou obra em termos de exatidão, considerando com que ela foi feita (precisão interna).
B - Julgar um documento ou obra em termos da consistência do argumento; relações entre suposições, evidência e conclusões (consistência interna).
C - Reconhecer os valores e pontos de vista usados numa obra (critérios internos).
D - Julgar uma obra comparando-a com outra obra relevante (critérios externos).
E - Julgar uma obra usando um conjunto de critérios ou padrões dados por outrem (critérios externos).
F - Julgar uma obra usando um conjunto de critérios ou padrões produzidos explicitamente pelo sujeito mesmo (critérios externos).

Ações típicas destas habilidades são: avaliar, comparar, contrastar, concluir, criticar, defender, justificar, interpretar, dar suporte, validar.

3 - Testando Avaliação A:

Julgar precisão de uma obra em seus detalhes. Exemplo:

Dois pesquisadores fizeram o seguinte experimento sobre a influência da ansiedade sobre o desempenho escolar.

O pesquisador A criou experimentalmente dois grupos de sujeitos, um ansioso e o outro não-ansioso, e verificou que os sujeitos ansiosos apresentavam desempenho mais fraco que os não-ansiosos.

O pesquisador B criou experimentalmente três grupos de ansiedade: muito ansiosos, moderadamente ansiosos, não ansiosos. Além disso, mediu o nível intelectual de todos estes sujeitos. Verificou que a ansiedade moderada ajuda o desempenho escolar em sujeitos com níveis intelectuais mais altos.

Agora, explicite o que o pesquisador A não controlou e que foi controlado pelo pesquisador B nesta pesquisa.

4 - Testando Avaliação B:

Julgar a lógica do documento, isto é, como as partes e detalhes se organizam e se relacionam de modo a formar um todo coerente. Exemplo:

A seguir você encontrará uma série de pares de argumentos tratando de um só assunto. Para cada par, você deverá localizar a fonte de sua oposição, real ou aparente. Para tanto considere o contexto donde a citação foi retirada. Utilize a escala A a E para responder.

1) *Par de argumentos*:
   a) "A descoberta científica depende sempre de um pensamento feliz, cuja origem não pode ser traçada; de um lance fortuito do intelecto, que se situa acima de todas as regras".
   b) "Descobrir as ciências vai além dos limites do engeniosidade do homem e deixa pouco campo para a excelência individual".

   Escala: A oposição entre estes dois argumentos

A - é puramente verbal.
B - se deve a que a) se refere à "descoberta científica" e b) ao "descobrir ciências".
C - se deve ao fato de que a) tem a haver com indução e b) com dedução.
D - se deve a duas diferentes suposições com referência à relação entre dados e idéias.
E - se deve a duas diferentes suposições com referência à relação entre observação e experimento.

2) *Par de argumentos*:
   a) "seres humanos em sociedade não possuem outras propriedades senão as que se podem derivar e ser reduzidas às leis da natureza dos indivíduos".
   b) "Assim, os grandes movimentos de entusiasmo, indignação e piedade numa multidão não se originam senão das consciências individuais".

   Escala: A oposição entre os dois argumentos

   A - é meramente verbal.
   B - implica concepções opostas da psicologia do indivíduo humano.
   C - é atribuível à inclusão de 'fenômenos sociais' citados em b) dentro do escopo das 'leis da natureza do indivíduo humano' citado em a).
   D - é atribuível à diferença entre a 'natureza' de a) e a 'consciência' de b).
   E - expressa atitudes opostas com respeito à concepção de Mill sobre o método 'químico' nas ciências sociais.

5 - Testando Avaliação C:

Identificar (sem julgar) valores, pontos de vista e pressupostos que o autor usa na obra. Exemplo:

É feita freqüentemente a afirmação de que "Física e química são ciências básicas; astronomia e geologia são ciências derivadas". Qual das seguintes constitui a melhor interpretação desta afirmação?

A. Física e química se baseiam em fundamentos sólidos de leis demonstradas; muita coisa em astronomia e geologia constitui pura especulação.
B. O desenvolvimento da astronomia e da geologia necessita o uso da física e da química; o converso não sendo verdadeiro.
C. Toda a matéria da astronomia e da geologia poderia ter sido derivada através do uso das leis e métodos da física e da química.
D. Física e química são de importância mais fundamental para a atividade do homem do que a astronomia e a geologia.
E. É possível se realizarem experimentos em laboratório no caso da física e da química, mas não no caso da astronomia e da geologia.

6 - Testando Avaliação D e E:

Julgar uma obra usando critérios externos, a saber, comparando-a com outra obra relevante (D) ou com um conjunto de critérios ou padrões previamente dados (E). Exemplo:

- As questões [1 e 2] tratam da seguinte passagem referente à natureza geral da ciência
- "No fundo, a ciência admite apenas um teste da validade de suas teorias, a saber, *concordância*: conseqüências tiradas puramente por raciocínio dedutivo a partir de observações não podem contradizer teorias da ciência.
- O objetivo da ciência consiste em descobrir ordem no mundo, isto é, descobrir relações que unem os vários fenômenos observados. Para este propósito é necessário construir teorias. Estas teorias sempre postulam a existência de entidades (tais como o átomo, campos de força ou massa do sol) que não podem ser diretamente percebidas em qualquer observação, mas que servem para unificar nossa imagem do mundo que está atrás das nossas observações. A unificação consiste no fato de que, por meio de entidades não observáveis e das relações postuladas entre elas, nós podemos tirar conclusões de um conjunto de observações sobre outro conjunto similar.
- Em particular, as teorias científicas se apresentam freqüentemente em forma quantitativa, as entidades postuladas são caracterizadas por números, e as relações entre elas são expressas matematicamente. Neste caso, as quantidades "teóricas" são passíveis de computação com base na observação, sendo que as observações elas mesmas devem, por isso, ser quantitativas, isto é, elas devem ser dados de medida. Ademais, cada quantidade "teórica" deve, pelo menos em algumas situações, poder ser computável independentemente a partir de *mais de um conjunto* de medidas. O acordo entre as medidas é a concordância que testa o teoria; se tais determinações independentes não forem possíveis, não é possível se testarem os resultados da teoria; a quantidade teórica não teria qualquer utilidade e seria apenas redundante".

1) Quais das observações seguintes, tomadas por si sós, fornece um teste da "concordância" obtida por Newton na teoria das partículas da luz?
   A. Observação da trajetória de uma partícula, dentro das condições descritas pelo primeiro teorema de Newton referente aos "corpos muito pequenos".
   B. Medida do ângulo de incidência e o ângulo de refração de um único raio de luz atravessando de um meio a outro.
   C. Medida do ângulo de incidência e do ângulo de reflexão de um único raio de luz refletido de uma superfície que separado dois meios diferentes.
   D. Medida do ângulo no qual ocorre reflexão total para um único raio de luz passando da água para o ar.
2) Qual das seguintes afirmações seria considerada, pelo autor da passagem acima citada, uma razão decisiva para rejeitar uma teoria científica?
   A. A teoria não é quantitativa.
   B. A teoria não é concordante com certos fenômenos.
   C. A teoria contém elementos redundantes.
   D. A teoria contém postulados que atribuem a certas entidades propriedades que diferem daquilo que se tem observado em objetos ou processos visíveis.
   E. Cada qual destas afirmações constituiria uma razão decisiva para rejeitar a teoria.

7 - Testando Avaliação F:

Julgar uma obra usando critérios explicitamente estabelecidos pelo próprio aluno a priori. Exemplo:

Escreva uma redação de 400 a 600 palavras sobre um dos quatro poemas designados para estudo de casa. A redação deve ter em conta as três condições seguintes:

A - Deve dar um ou mais julgamentos sobre o poema (por ex., o valor de sua intenção, o sucesso dos meios para os fins, sua veracidade, beleza, etc.)
B - Deve tornar explicita a natureza de cada critério utilizado e explicar que suposições referentes à natureza e aos meios da poesia dão a este critério uma base significativa para o julgamento.
C - Deve discutir em detalhes que partes ou aspectos do poema são pertinentes ao julgamento dado - como prova de que suas conclusões são justificadas.

G - Técnicas de Avaliação dos Objetivos de Domínio Afetivo

1 - Valores a Avaliar

Os objetivos afetivos a serem avaliados no contexto do ensino variam segundo os interesses do avaliador. Bloom e colaboradores (1971), como vimos acima, têm toda uma lista destes objetivos. Henerson, Morris e Fitz-Gibbon (1978) apresentam um elenco de categorias de objetivos útil para o presente contexto. Os valores que os objetivos afetivos normalmente procuram medir se caracterizam por atitudes que se espera o aluno desenvolver no seu treinamento escolar. Henerson e cols. (1978) sumariam estas atitudes em

- atitudes para consigo mesmo
  - auto-estima
  - auto-percepção
  - auto-conceito
  - auto-realização
  - integração da personalidade
  - força do ego
  - autoconfiança
  - locus de controle
- atitudes para com a escola e assuntos a ela referentes
  - assuntos específicos da escola
  - colegas de escola
  - professores
  - ambiente da escola
  - processo de aprendizagem
  - educação em geral
- atitudes para com os outros
  - confiança nos outros
  - aceitação dos outros
  - preocupação com os outros

. estratégias no trato com os outros
. "insight" social
. filiação a um grupo
. raça, cultura, religião dos outros

– atitudes para com o trabalho e interesses gerais.

Outros autores (como Mager, 1968) e cada avaliador podem desenvolver o elenco de atitudes que acharem pertinentes no contexto da escola. Parece que qualquer atitude que os psicólogos estudam ou estudaram pode entrar neste elenco. O importante é justificar, para o contexto da escola, a pertinência dos valores incluídos para a avaliação. Por exemplo, parece estranho que no elenco dado acima não entrem valores de personalidade que podem ter impacto relevante na aprendizagem, como por exemplo, a motivação, a ansiedade, etc. Enfim, esta área do domínio afetivo está ainda à mercê dos interesses do avaliador, mais que haver uma taxionomia que faça sentido para todos.

2 - Técnicas de Avaliação do Domínio Afetivo

Existem nesta área da avaliação afetiva as técnicas mais variadas que vão desde a observação do comportamento até os testes psicológicos. Grande parte destas técnicas estão desenvolvidas em outros capítulos deste livro e, portanto, serão aqui apenas elencadas. Particularmente relevante é o capítulo sobre a construção de testes referentes a construto (cap. 3). Mas as técnicas possíveis são muitas, as quais podem ser brevemente assinaladas como segue:

– Testes psicológicos, inventários de personalidade (cap. 3)
– Diferencial semântico (cap. 6)
– Escalas de atitude (cap. 5 e cap. 4)
– Observação do comportamento
– Questionários e "surveys" (cap. 10)
– Testes sociométricos
– Entrevistas
– Portfólios

Bibliografia

Anastasi, A. (1988). *Psychological testing*. Sixth edition. New York: Macmillan Publ. Co.

Binet, A. & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique, 11*, 191-244.

Bloom, B.S. (Ed. - 1956). *Taxonomy of educational objectives: The classification of educational goals*. Handbook 1. *Cognitive domain*. New York: McKay.

Bloom. B.S., Hastings, J.T., & Madaus, G.F. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill Book Co.

Crooks, T.J. (1988). The impact of classroom evaluation practices on studenst. *Review of Educational Research, 58*, 438-481.

Davies, I.K. (1982). *O planejamento de currículo e seus objetivos*. Ed. Saraiva.

French, W. & Associates (1957). *Behavioral goals of general education in high school*. New York: Russell Sage Foundation.

Gagné, R.M. (1965). The analysis of instructional objectives for the design of instruction. In R. Glaser (Ed.), *Teaching machines and programmed learning*. Vol. 2. Data and directions. Washington, DC: National Education Association, 21-65.

Gagné, R.M. (1963). Learning and proficiency in mathematics. *Mathematics Teacher, 56*, 623.

Gagné, R.M. (1980). *Princípios essenciais da aprendizagem para o ensino*. Porto Alegre: Ed. Globo.

Goleman, D. (1996). *Inteligência emocional*. Rio de Janeiro, RJ: Objetiva.

Goldshalk, F.I., Swineford, F., & Coffman, W.E. (1966). *The measurement of writing ability*. New York: College Entrance Examination Board.

Green, S.B., Halpin, G., & Halpin, G.W. (1990). *The emphasis on rote memory items on classroom testes: Why are teachers so interested in hearing their own lectures*. Paper presented at the annual meeting of the National Council on Measurement in Education. Boston.

Greeno, J.G. (1980). Some examples of cognitive task analysis with instructional implications. In R.E. Snow, P.-A. Federico, & W.E. Montague (Eds.), *Aptitude, learning, and instruction. Vol. 2: Cognitive process analyses of learning and problem solving*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gronlund, N.E. (1974). *Elaboración de tests de aprovechamiento*. México: Editorial Trillas.

Haladyna, T.M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, NJ: Lawrence Erlbaum Associates, Pubs.

Harrow, A.J. (1972). *A taxonomy of the psychomotor domain*. New York: David McKay Co.

Henerson, M.E.; Morris, L.L. & Fitz-Gibbon, C.T. (1978). *How to measume attitudes. Chapter 4: Finding an Existing Measure*. Beverlly Hills, California: SAGE Publications, Inc.

Henry, N.B. (Ed. - 1946). *The measurement of understanding: The forty-fifth yearbook of the National Society for the Study of Education*. Chicago: Univesity of Chicago Press.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). *Item response theory: Applications to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

Kelley, T.L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30*, 17-24.

Krathwohl, D.R., Bloom, B.S., & Masia, B.B. (1964). In B.S. Bloom (Ed.), *Taxonomy of educational objectives: The classification of educational goals*. Handbook 2. *Affective domain*. New York: McKay, 176-185.

Kubiszyn, T. & Borich, G. (1984 - 3rd ed.). *Educational testing and measurement*. Glenview, IL: Scott, Foresman/Little, Brown Higher Education.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Mager, R.F. (1968; trad. 2a. ed. 1979). *Atitudes favoráveis ao ensino*. Porto Alegre: Editora Globo.

Mager, R.F. (1981). *Medindo os objetivos de ensino* ou "conseguiu um par adequado". Porto Alegre: Editora Globo.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3d. ed.). New York: Macmillan, 13-103.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Piaget, A. (1952). *The origins of intelligence in children*. New York: International Universities Press.

Pinard, A. & Laurendeau, M. (1964). A scale of mental development based on the theory of Piaget: Description of a project. *Journal of Research in Science Teaching, 2*, 253-260.

Roid, G.K. & Haladyna, T.M. (1982). *Toward a technology of test-item writing*. New York: Academic Press.

Shavelson, R.J., Baxter, G.P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*, 22-27.

Snow, R.E. & Lohman, D.F. (1989). Implications of cognitive psychology for educational measurement. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education and Macmillan, 263-332.

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

Sternberg, R.J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. Hillsdale, NJ: Erlbaum.

Sternberg, R.J. (1979). The nature of mental abilities. *American Psychologist, 34*, 214-230.

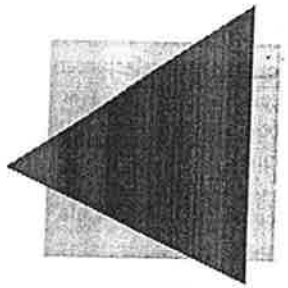
Stiggins, R.J., Griswold, M.M., & Wikelund, K.R. (1989). Measuring thinking skills through classroom assessment. *Journal of Educational Measurement*, 26, 233-246.

Telebrás. (1989). *Manual de palnejamento instrucional da Telebrás*. Brasília, DF: Telebrás.

Tyler, R.W. (1950). *Basic principles of curriculum and instruction*. Chicago, IL: University of Chicago Press.

Tyler, R.W. (1954). *The fact-finding study of the testing program of the United States Armed Forces Institute, 1952-1954*. Report to the USAFI, University of Chicago.

Wiggins, G. (1989). Teaching to the (authentic) test. *Educational Leadership, 76*, 41-47.



ANEXO

OUTRAS TAXIONOMIAS EDUCACIONAIS

Objetivos Segundo Mager

Para este teórico da aprendizagem, o objetivo descreve o resultado desejado no curso (Mager, 1975). Representa o que o aluno poderá fazer depois da experiência de aprendizagem. Os objetivos, quando definidos claramente, permitem que o aprendiz avalie o seu progresso e reoriente o seu desempenho de forma tal que aumente a sua probabilidade de sucesso.

Mager (1962) identifica três componentes:

1 - Comportamento: "Primeiro, identifique o comportamento final pelo nome; você pode especificar o tipo de comportamento que será aceito como evidência de que o aluno alcançou o objetivo".

2 - Condições: "Segundo, tente definir o comportamento desejado ainda mais, descrevendo as condições importantes sob as quais se espera que o comportamento ocorra".

3 - Padrões: "Terceiro, especifique o critério de desempenho aceitável, descrevendo como o aluno deve se empenhar para ser considerado aceitável".

Um objetivo com as três dimensões, segundo Mager, salvaguarda a orientação. Não significa que todo objetivo precisa incluir os três componentes. Se o objetivo comunica com sucesso a intenção do professor, os detalhes excessivos são desnecessários.

Quando se define o componente comportamento de um objetivo, é importante que se defina claramente o que se espera que o aluno faça quando tiver que demonstrar competência. Assim, o uso do verbo de ação é o ponto sobre o qual a declaração do comportamento pode ser construída.

Com a definição do componente comportamento, os outros dois, condições e padrões, podem ser construídos, tendo sempre por referência o próprio componente comportamental.

O componente condições de um objetivo é o passo seguinte na montagem de um objetivo específico. Geralmente, implica na especificação das condições ou limitações que serão atribuídas ao aluno no instante de evidenciar o que realizou.

"Algumas vezes nos exames pede-se aos alunos para descreverem como eles usariam a tábua de logaritmos e as réguas de cálculo para resolver um determinado problema, quando eles estão só acostumados a usá-las mecanicamente. Estudantes quer foram ensinados a usar o Atlas em Geografia, fazem exames sem os mes-

mos. O mesmo acontece com dicionários, tabelas, fórmulas, material de referência, etc." (Manual de Planejamento Instrucional da TELEBRÁS, 1989).

As condições podem ser das mais diversas ordens e se relacionam por exemplo com a extensão dos problemas que os alunos precisam aprender a resolver. As condições ambientais também devem ser definidas. Pode-se resolver o problema em casa, na biblioteca da escola ou ainda no laboratório. Finalmente, algumas condições envolvem exigências físicas do aluno. Por exemplo, deitar, subir escadas, agachar, que pode aumentar a dificuldade de uma tarefa e o tempo necessário à sua execução.

Para Mager, o elemento final de um objetivo é o componente padrões. Especifica os padrões que o aluno deve alcançar para que se considere se existe domínio e realização. São três basicamente os padrões envolvidos. O primeiro corresponde à porcentagem de problemas que precisam ser desempenhados com êxito. Os alunos devem acertar todos ou alguns apenas. O segundo padrão fala das tolerâncias sob as quais os estudantes devem trabalhar. Por exemplo, eles têm que dar a resposta a todas as capitais ou apenas de dez? Por último, o padrão de tempo. Admite-se, por exemplo, que as capitais possam ser citadas num determinado espaço de tempo.

Estas três dimensões dos objetivos são complementares e se encaixam tal qual um quebra-cabeça (veja figura 7-6).

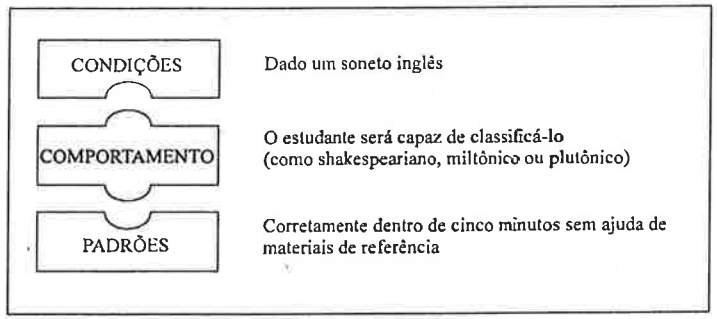


Fig. 7-6. Componentes de um objetivo instrucional, segundo Mager

É importante lembrar que os objetivos precisam ser combinados para formar uma progressão ou unidade de ensino, que combinadas formam um programa ou um curso. Todo este procedimento faz parte do planejamento sistêmico de ensino. Esquematizando Mager, temos:

Desempenho	ação observável que o aluno deverá evidenciar
Condição	situação sob a qual o comportamento será demonstrado
Padrão	critério satisfatório de desempenho

Alguns objetivos, definidos segundo o modelo sugerido por Mager, são apresentados no quadro a seguir:

1. Modelo
   - (Condição) Dada uma reprodução da pintura de David "The Oath of the Horatii", o estudante definirá seis razões que possam ser usadas para substanciar o argumento de que esta pintura representou um novo tipo de classicismo revolucionário na arte.
   - (Padrões) Deve-se fazer referência à composição, ação, expressão e acessórios do significado da pintura.
2. Exemplos
   - Dados os detalhes das estatísticas climáticas (temperatura mensal, chuva e umidade) de dez cidades, o estudante identificará corretamente por escrito o nome do tipo de clima associado a cada conjunto de estatísticas. Precisa ser dado um mínimo de três razões substanciando o raciocínio, atrás de cada uma das dez decisões. O limite mais baixo de atuação aceitável será 8 identificações corretas entre 10, dentro de um período de 30 minutos.
   - Dado um esqueleto humano, o estudante identificará oralmente e corretamente o nome de todos os ossos da mão esquerda e da perna direita abaixo do joelho. Não haverá punição em caso de adivinhação. A tarefa precisa ser completada em 30 minutos, sem o auxílio de material de referência.
   - Dada uma cópia da biografia de Cromwell "The Lord Protector", por Antonia Fraser, o estudante lerá o livro e escreverá uma dissertação de dez páginas, relatando o material na "Parte UM: O controle de si mesmo", lidando com os anos de 1599 até 1642, e relatando os fatos da " Parte Três" ou da "Parte Quatro". A tarefa deverá ser completada por volta de 22 de agosto. O material da dissertação deve procurar justificar o comentário de John Milton que diz que Cromwell, "primeiro adquiriu o controle de si mesmo e conquistou a maioria dos sinais de vitória através de si mesmo, para que, no primeiro dia de batalha contra o inimigo, ele já fosse um veterano em armas, acostumado às práticas nas armadilhas e exigências da guerra".
   - O estudante oralmente definirá os termos: carga elétrica, corrente elétrica, força eletromotora, diferença de resistência e diferença de potência. Estes cinco conceitos devem então ser usados para explicar o princípio da Lei de Ohm.
   - O estudante construirá um trabalho reticular simples contendo até dez fatos, numa narrativa descrevendo as atividades, sem erro e sem material de referência, dentro de um período de tempo que não exceda trinta minutos.

Objetivos Segundo Gagné

Gagné fez um resumo dos trabalhos de Mager e Miller (Robert Miller, empregado da IBM, apresentou um trabalho sobre formulação de objetivos, muito considerado no contexto educacional em geral. Embora seu interesse se concentrasse na especificação de habilidades físicas complexas, a sua proposta tem sido para

habilidades cognitivas e psicomotoras). Identificou, primeiro, as condições ("dadas duas dezenas ligadas pelo sinal +"), a seguir o comportamento (nesse caso seria "expõe"), o objeto a ser trabalhado ("por escrito"), e por último os padrões que denotam sua precisão ("o nome do número que representa a soma das duas dezenas").

Este esquema representa a estrutura estímulo-resposta de Gagné. O seu trabalho teve profunda influência na ciência do ensino com seus respectivos materiais curriculares.

Gagné é mais detalhista que Mager e Miller e enfatiza a importância de se definir um objetivo de maneira explícita. Insiste na importância em se definir o que o aluno estará fazendo ao final do curso; ressaltou, contudo, que as metas que estão muito distantes em termos de tempo devem ser evitadas.

Com a finalidade de operacionalizar ainda mais um objetivo de ensino, Gagné insiste na distinção entre "verbos de ação" e "habilidades aprendidas". Para ele, os verbos de ação não são os componentes mais importantes do objetivo, mas sim o verbo que descreve a habilidade exigida.

Par Gagné, existe uma relação de alguns verbos que de fato representam as habilidades humanas: discrimina, identifica, classifica, demonstra, gera, origina, expõe, executa, escolhe. Num objetivo envolvendo alunos "testando uma hipótese", a habilidade talvez se centralize no verbo "discrimina" (Manual de Planejamento Instrucional da TELEBRÁS, 1989).

São sugeridos cinco verbos de habilidades humanas que envolvem habilidades intelectuais. Usar estas cinco palavras significa preservar as distinções do exemplo acima. O quadro a seguir exemplifica um objetivo bem definido:

1. Modelo
   - (Condição) Dada uma reprodução da pintura de David "The Oath of the Horatii" (habilidade), o estudante gera (objeto) seis razões (ações) que podem ser usadas para substanciar o argumento de que esta pintura representou um novo tipo de revolução no classicismo da arte. (Restrição) Nas razões deve ser feita referência à composição, ação, expressão e acessórios.
2. Exemplos
   - Dada uma lista de dez sentenças complexas, identificar as frases, classificando-as como adjetivas ou adverbiais, de acordo com a definição previamente dada, sem material de referência ou outro tipo de ajuda.
   - Dado um conjunto científica Norstedt, o estudante demonstra o relacionamento entre a diferença potencial, resistência e corrente, montando uma situação prática que permita que a relação seja observada e os dados recolhidos.
   - Dado o radiador de um carro para ser enchido com a quantidade apropriada de aditivo anticongelante para um inverno britânico, o aluno identifica a quantidade de anticongelante requerida para temperaturas que não vão além de -10°C, usando um hidrômetro e tabela de cálculo anticongelante, executa a tarefa com a precisão aproximada de 1/8 de litro.

- Dada a responsabilidade de formular objetivos específicos para um curso de ciências políticas, o estudante-professor identifica os propósitos, metas e objetivos específicos da informação dada e coloca as mesmas em forma de lista, usando o formato sugerido pelo Dr. Robert Gagné sempre que for apropriado.

Um objetivo definido no formato proposto por Gagné apresenta cinco componentes. Dois a mais do que no modelo de Mager. São eles:

- Ação: demonstra um comportamento descrito por um verbo de ação;
- Objeto: determina o que tem de ser produzido ou processado;
- Situação: é o momento ou situação com a qual o aluno se confronta quando lhe é solicitado que evidencie algum desempenho:
- Acessórios e outras restrições: como será conduzido o desempenho?
- Habilidades a serem aprendidas: é o tipo de desempenho que se espera do aluno.

Objetivos Instrucionais redigidos conforme esse esquema denotam a "natureza refinada do processo instrutivo" (Manual de Planejamento Instrucional da TELEBRÁS, 1989). Assim, as habilidades que eles representam constituem uma taxionomia que é útil na construção dos testes de aprendizagem. A seguir, a síntese dos componentes de um objetivo instrucional segundo Gagné:

Ação	Descreve o comportamento por meio de um verbo de ação
Objeto	Descreve o que será produzido ou processado
Situação	é a situação com que o aluno se confronta quando lhe é solicitado que faça alguma coisa
Ferramentas e Restrições/Acessórios	Relacionando-se a como o desempenho precisa ser conduzido
Desempenho ou Habilidades a serem aprendidas	referem-se ao tipo de habilidade que se espera que o aluno demonstre

Do ponto de vista dos tipos de objetivos possíveis, fazem-se algumas classificações, a saber:

- Objetivo Geral: Representa o conjunto de desempenhos que se pretendem que os alunos alcancem ao término do curso. Eles deverão, ao final da disciplina, ter adquirido as competências definidas e englobadas pelo objetivo geral. Os conhecimentos e habilidades abrangidos pelo objetivo geral representam um somatório daqueles referentes aos objetivos específicos. Portanto, estes, em seu conjunto, devem representar o objetivo geral do curso, do qual são desdobramentos. Os objetivos gerais dão uma orientação em direção à qual se quer que a mudança de comportamento dos educandos ocorra.
- Objetivo Específico: Representa o conjunto de conhecimentos e habilidades básicas que os alunos deverão alcançar. Os objetivos específicos podem ser expressos em termos abstratos ou em comportamentos. Trata-se realmente das definições dos objetivos que podem ser constitutivas ou operacionais. Aquelas definem ou expressam os objetivos em termos de abstrações, ao passo que estas os expressam em termos de comportamentos. Por exemplo, se meu objetivo for

"aumentar no aluno a compreensão do sistema numérico", este é um objetivo abstrato, porque como poderá o aluno mostrar que compreendeu? Ao passo que "demonstrar o teorema de Pitágoras é um objetivo comportamental ou operacionalizado, pois o aluno deve saber demonstrar o teorema" (Mager, 1981).

- Objetivo Intermediário: Representa o desdobramento de um objetivo específico mais complexo. Para se definir um objetivo intermediário, realiza-se a análise de cada objetivo específico. Os conhecimentos necessários para o seu alcance irão derivar um objetivo intermediário.

Observe o exemplo a seguir:

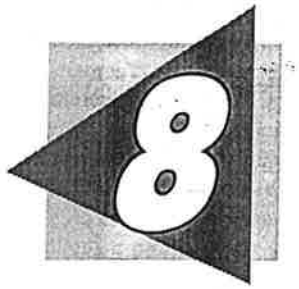
Objetivo Geral: Ao final do curso o aluno deverá ser capaz de relacionar os objetivos para um dado curso.

Objetivo específico	Conhecimento necessário	Objetivo intermediário
Redigir os objetivos de pré-requisito para um dado curso	- conceito pré-requisito	- conceituar pré-requisito
	- componentes dos objetivos	- descrever os componentes de um objetivo de pré-requisito
	- critérios para definição do pré-requisito	- identificar critérios para delimitação de pré-requisitos para um curso

O esquema para classificação dos objetivos instrucionais pode ser o seguinte (quanto ao nível de especificação):

- Geral: descreve as intenções a longo prazo;
- Específico: descreve os conhecimentos, habilidades e atitudes esperadas após determinadas práticas instrucionais;
- Intermediário: descreve os conhecimentos ou habilidades necessárias ao alcance de um objetivo mais complexo.

Importante: É necessário registrar a relação entre os objetivos gerais e os específicos. Tanto os objetivos gerais entre si, como os específicos relacionados aos gerais, devem ser complementares. "Uma plataforma de objetivos gerais é ainda um guia inadequado para os aspectos específicos da instrução... A especificação tem o propósito de ajustar os objetivos gerais ao conteúdo específico e às necessidades de desenvolvimento do grupo em particular".



Testes Centrados em Critério (CRT)

Leandro S. Almeida

1 - Introdução

A designação *"criterion-referenced tests" (CRT)* foi usada pela primeira vez nos meios acadêmicos por Glaser (1963). A sua popularidade não foi imediata, podendo-se apontar um "período de gestação" até 1970 (Popham, 1978). A edição de uma obra conjunta em 1971 lançou definitivamente o tema (Popham, 1971). Nos finais dos anos 70 não só estavam publicados mais de meio milhar de artigos na área (Hambleton, Powell & Eignor, 1979), como alguma confusão se tinha instalado no seu seio.

Os principais aspectos dessa confusão tinham a ver com a definição deste tipo de provas e com as diversas expressões que eram usadas para as designar. Ambos os aspectos estão interligados. Por exemplo, nos finais dos anos 70, Gray (1978) menciona mais de 50 definições diferentes, ao mesmo tempo que Linn (1981) menciona a grande proximidade entre os testes agrupados nas seguintes designações "criterion-referenced tests", "domain-referenced tests", "curriculum-related tests", "objectives-referenced tests", ou "competency and mastery tests". Estes conceitos salientam, ou mais o domínio a especificar para o teste, a natureza dos itens ou a interpretação dada ao resultado. Talvez o termo *"criterion-referenced tests"* tenha o mérito de juntar os vários aspectos: determinação do *nível* do sujeito num *domínio* de comportamento bem delimitado (Popham, 1975), assegurando inferências válidas a partir do desempenho do sujeito na amostra representativa de itens que integram o teste (Hambleton, 1982).

Os dois parâmetros que melhor definem os "testes centrados em critério" são, como teremos ocasião de ilustrar ao longo deste capítulo, o domínio delimitado de competências em que incide a avaliação e a existência de um nível prévio definindo o desempenho satisfatório e não satisfatório. Eles encontram-se presentes na definição mais freqüentemente aceite: *"Um teste referenciado a critério é aquele que foi deliberadamente construído para produzir medidas que são diretamente interpretáveis em termos de padrões específicos de desempenho ("A criterion-referenced test is one that is deliberately constructed to yield measurements that are directely interpretable in terms of specified performance standards"*, Glaser & Nitko, 1971, p. 653). Este último ponto diferencia claramente este grupo de testes

dos testes clássicos, ditos "norm-referenced tests" (NRT), cujos resultados são interpretados com base em "normas" obtidas em amostras representativas de indivíduos.

Na tabela 8-1 apresenta-se um exemplo de CRT, no caso concreto procurando avaliar a capacidade de segmentação de sons iniciais das palavras por parte da criança, reduzida à segmentação das consoantes em que vai incidir um primeiro programa de treino de identificação de sons consonânticos no jardim de infância. O critério de sucesso (mestria) integra as instruções dadas ao avaliador.

Tabela 8-1. Segmentação fonética de palavras

Domínio: Pronunciar palavras que começam com o mesmo som consonântico.	
Mestria: A criança tem que pronunciar corretamente 2 das 3 palavras que começam com, pelo menos, 8 das 10 consoantes.	
REGISTO (Nº das palavras pronunciadas corretamente)	
t (toca, tua, tapa) __	v (vela, vão, vinha) __
p (pata, pele, pia) __	r (rela, roca, rato) __
n (nota, nada, nova) __	c (cão, cola, capa) __
l (lata, lua, linha) __	s (selo, soca, sapo) __
d (dela, doca, dado) __	j (jarro, jogo, juiz) __
Nº das consoantes com notas de 2 ou 3: _____	
Mestria? Sim ____ Não ____	

2 - Utilização dos CRT

Historicamente o uso dos CRT centrou-se na educação. A sua utilização pelos psicólogos escolares pretendia assegurar uma informação mais individual de cada aluno e a ligação dessa informação à programação das atividades educativo-escolares. Algumas críticas aos testes de inteligência acentuavam a sua dificuldade em informar do potencial de aprendizagem, das formas individuais como essa aprendizagem se processava e dos níveis atingidos no final de cada unidade ou período escolar. Este tipo de informações era decisivo quando as práticas de ensino dos professores nos Estados Unidos da América se encontravam fortemente influenciadas pelo movimento da aprendizagem programada ou da aprendizagem por objetivos.

Em primeiro lugar, a existência de um critério absoluto previamente fixado que o sujeito atinge ou não - ter ou não mestria no comportamento em causa - permitia estimar a aprendizagem realizada pelo aluno e, indiretamente, a eficácia do ensino do professor. Em segundo lugar, a definição prévia do critério de mestria na aprendizagem assegurava uma apreciação direta do resultado individual do aluno (por exemplo, conseguir estar atento nos últimos dez minutos da aula, conhecer o significado de 10 em 12 palavras), não se tornando necessário comparar o desempenho do aluno em relação aos seus pares. Aliás, a não necessidade de referência aos níveis de realização dos outros colegas para apreciação dos resultados individuais constitui uma das principais distinções entre os testes centrados no *critério* e os testes centrados na *norma.* Ao critério absoluto fixado nos CRT contrapõem-se os padrões relativos às amostras (normas) nos testes tradicionais (Gronlund, 1973).

Em terceiro lugar, tanto a mestria como a não mestria num comportamento após um dado período de aprendizagem constituem informação relevante para os gestores do ensino-aprendizagem. Várias decisões ficam, a partir daí, legitimadas. A não mestria, por exemplo, pode levar a questionar o critério fixado (nível de exigência inapropriado), a inventariar necessidades suplementares de aprendizagem por parte dos alunos para a conveniente resolução da tarefa ou a sugerir metodologias de ensino do professor mais ajustadas. A tomada desta informação ao longo da escolaridade do aluno *(portfólio)* pode constituir, ainda, material relevante para as decisões escolares e profissionais a tomar pelos alunos e seus encarregados de educação ao longo da escolaridade.

Em quarto lugar, uma das razões para a popularidade dos CRT no ensino prende-se com a progressiva consciência dos educadores (e dos movimentos de reforma educativa) da importância do ensino individualizado. Tal prática exige um sistema de avaliação freqüente e centrada no processo e nas aquisições da aprendizagem. Situação idêntica é defensável na educação especial. Importa que a avaliação sirva mais de perto o processo de ensino-aprendizagem, permitindo (i) a monitorização dos alunos e dos professores quanto aos progressos escolares, (ii) a definição dos objetivos para uma determinada intervenção, (iii) o diagnóstico de deficiências na aprendizagem, (iv) a avaliação de programas de intervenção sócio-educativa ou (v) a certificação das aprendizagens conseguidas pelos alunos.

Em quinto lugar, os CRT parecem muito mais adequados aos objetivos, em geral, de toda a avaliação escolar. Os CRT estão construídos já com a hipótese subjacente do crescimento na experiência ou nos conhecimentos, e daí mais orientados para avaliação dos ganhos progressivos intra-individuais - objetivo edumétrico - do que para a avaliação das diferenças estáveis entre os indivíduos - objetivo psicométrico (Carver, 1974). Gronlund (1973, ps. 42-45) refere várias situações de tal utilização: (i) avaliação inicial ou de diagnóstico (avaliar pré-requesitos necessários para a aprendizagem, determinar o lugar de cada aluno numa seqüência de aprendizagens, determinar a porção de uma área de instrução que os alunos já dominam, fornecer uma base para avaliação da aprendizagem adquirida durante a instrução, e sugerir formas concretas de superação de deficiências diagnosticadas); (ii) avaliação formativa ou contínua (determinar que objetivos de uma unidade foram atingidos, detectar deficiências com base nos itens falhados, prescrever as medidas julgadas convenientes para superar as dificuldades de cada aluno, e usar a informação para a promoção da própria instrução em termos de métodos, estratégias ou seqüências de aprendizagens); e (iii) avaliação sumativa (geralmente confinada ao final de um programa de formação e tendo essencialmente em vista a certificação dos indivíduos com bases nos produtos finais atingidos).

O uso dos CRT na avaliação sumativa ou de certificação final dos alunos requer, na maioria das vezes, que o nível de mestria fixado seja desdobrado e diferenciado. Se para a avaliação contínua era suficiente a fixação de um critério único de mestria, os objetivos de seriação e de classificação dos alunos no final de um ano escolar ou nível de escolaridade aconselham o recurso a provas diferenciadoras das aprendizagens e das competências desenvolvidas. Nestes casos, os testes centrados em critério poderão fixar vários níveis de realização ou de mestria. Por exemplo, podemos apontar um nível A ou um desempenho igual ou superior

"aumentar no aluno a compreensão do sistema numérico", este é um objetivo abstrato, porque como poderá o aluno mostrar que compreendeu? Ao passo que "demonstrar o teorema de Pitágoras é um objetivo comportamental ou operacionalizado, pois o aluno deve saber demonstrar o teorema" (Mager, 1981).

- Objetivo Intermediário: Representa o desdobramento de um objetivo específico mais complexo. Para se definir um objetivo intermediário, realiza-se a análise de cada objetivo específico. Os conhecimentos necessários para o seu alcance irão derivar um objetivo intermediário.

Observe o exemplo a seguir:

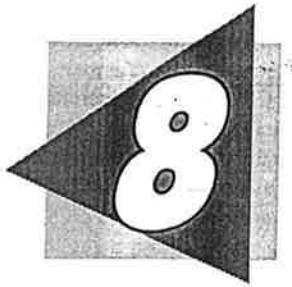
Objetivo Geral: Ao final do curso o aluno deverá ser capaz de relacionar os objetivos para um dado curso.

Objetivo específico	Conhecimento necessário	Objetivo intermediário
Redigir os objetivos de pré-requisito para um dado curso	- conceito pré-requisito	- conceituar pré-requisito
	- componentes dos objetivos	- descrever os componentes de um objetivo de pré-requisito
	- critérios para definição do pré-requisito	- identificar critérios para delimitação de pré-requisitos para um curso

O esquema para classificação dos objetivos instrucionais pode ser o seguinte (quanto ao nível de especificação):

- Geral: descreve as intenções a longo prazo;
- Específico: descreve os conhecimentos, habilidades e atitudes esperadas após determinadas práticas instrucionais;
- Intermediário: descreve os conhecimentos ou habilidades necessárias ao alcance de um objetivo mais complexo.

Importante: É necessário registrar a relação entre os objetivos gerais e os específicos. Tanto os objetivos gerais entre si, como os específicos relacionados aos gerais, devem ser complementares. "Uma plataforma de objetivos gerais é ainda um guia inadequado para os aspectos específicos da instrução... A especificação tem o propósito de ajustar os objetivos gerais ao conteúdo específico e às necessidades de desenvolvimento do grupo em particular".



Testes Centrados em Critério (CRT)

Leandro S. Almeida

1 - Introdução

A designação *"criterion-referenced tests" (CRT)* foi usada pela primeira vez nos meios acadêmicos por Glaser (1963). A sua popularidade não foi imediata, podendo-se apontar um "período de gestação" até 1970 (Popham, 1978). A edição de uma obra conjunta em 1971 lançou definitivamente o tema (Popham, 1971). Nos finais dos anos 70 não só estavam publicados mais de meio milhar de artigos na área (Hambleton, Powell & Eignor, 1979), como alguma confusão se tinha instalado no seu seio.

Os principais aspectos dessa confusão tinham a ver com a definição deste tipo de provas e com as diversas expressões que eram usadas para as designar. Ambos os aspectos estão interligados. Por exemplo, nos finais dos anos 70, Gray (1978) menciona mais de 50 definições diferentes, ao mesmo tempo que Linn (1981) menciona a grande proximidade entre os testes agrupados nas seguintes designações "criterion-referenced tests", "domain-referenced tests", "curriculum-related tests", "objectives-referenced tests", ou "competency and mastery tests". Estes conceitos salientam, ou mais o domínio a especificar para o teste, a natureza dos itens ou a interpretação dada ao resultado. Talvez o termo *"criterion-referenced tests"* tenha o mérito de juntar os vários aspectos: determinação do *nível* do sujeito num *domínio* de comportamento bem delimitado (Popham, 1975), assegurando inferências válidas a partir do desempenho do sujeito na amostra representativa de itens que integram o teste (Hambleton, 1982).

Os dois parâmetros que melhor definem os "testes centrados em critério" são, como teremos ocasião de ilustrar ao longo deste capítulo, o domínio delimitado de competências em que incide a avaliação e a existência de um nível prévio definindo o desempenho satisfatório e não satisfatório. Eles encontram-se presentes na definição mais freqüentemente aceite: *"Um teste referenciado a critério é aquele que foi deliberadamente construído para produzir medidas que são diretamente interpretáveis em termos de padrões específicos de desempenho ("A criterion-referenced test is one that is deliberately constructed to yield measurements that are directely interpretable in terms of specified performance standards"*, Glaser & Nitko, 1971, p. 653). Este último ponto diferencia claramente este grupo de testes

dos testes clássicos, ditos "norm-referenced tests" (NRT), cujos resultados são interpretados com base em "normas" obtidas em amostras representativas de indivíduos.

Na tabela 8-1 apresenta-se um exemplo de CRT, no caso concreto procurando avaliar a capacidade de segmentação de sons iniciais das palavras por parte da criança, reduzida à segmentação das consoantes em que vai incidir um primeiro programa de treino de identificação de sons consonânticos no jardim de infância. O critério de sucesso (mestria) integra as instruções dadas ao avaliador.

Tabela 8-1. Segmentação fonética de palavras

Domínio: Pronunciar palavras que começam com o mesmo som consonântico.
Mestria: A criança tem que pronunciar corretamente 2 das 3 palavras que começam com, pelo menos, 8 das 10 consoantes.
REGISTO (Nº das palavras pronunciadas corretamente)
t (toca, tua, tapa) __ — v (vela, vão, vinha) __
p (pata, pele, pia) __ — r (rela, roca, rato) __
n (nota, nada, nova) __ — c (cão, cola, capa) __
l (lata, lua, linha) __ — s (selo, soca, sapo) __
d (dela, doca, dado) __ — j (jarro, jogo, juiz) __
Nº das consoantes com notas de 2 ou 3: _____
Mestria? Sim ____ Não ____

2 - Utilização dos CRT

Historicamente o uso dos CRT centrou-se na educação. A sua utilização pelos psicólogos escolares pretendia assegurar uma informação mais individual de cada aluno e a ligação dessa informação à programação das atividades educativo-escolares. Algumas críticas aos testes de inteligência acentuavam a sua dificuldade em informar do potencial de aprendizagem, das formas individuais como essa aprendizagem se processava e dos níveis atingidos no final de cada unidade ou período escolar. Este tipo de informações era decisivo quando as práticas de ensino dos professores nos Estados Unidos da América se encontravam fortemente influenciadas pelo movimento da aprendizagem programada ou da aprendizagem por objetivos.

Em primeiro lugar, a existência de um critério absoluto previamente fixado que o sujeito atinge ou não - ter ou não mestria no comportamento em causa - permitia estimar a aprendizagem realizada pelo aluno e, indiretamente, a eficácia do ensino do professor. Em segundo lugar, a definição prévia do critério de mestria na aprendizagem assegurava uma apreciação direta do resultado individual do aluno (por exemplo, conseguir estar atento nos últimos dez minutos da aula, conhecer o significado de 10 em 12 palavras), não se tornando necessário comparar o desempenho do aluno em relação aos seus pares. Aliás, a não necessidade de referência aos níveis de realização dos outros colegas para apreciação dos resultados individuais constitui uma das principais distinções entre os testes centrados no *critério* e os testes centrados na *norma*. Ao critério absoluto fixado nos CRT contrapõem-se os padrões relativos às amostras (normas) nos testes tradicionais (Gronlund, 1973).

Em terceiro lugar, tanto a mestria como a não mestria num comportamento após um dado período de aprendizagem constituem informação relevante para os gestores do ensino-aprendizagem. Várias decisões ficam, a partir daí, legitimadas. A não mestria, por exemplo, pode levar a questionar o critério fixado (nível de exigência inapropriado), a inventariar necessidades suplementares de aprendizagem por parte dos alunos para a conveniente resolução da tarefa ou a sugerir metodologias de ensino do professor mais ajustadas. A tomada desta informação ao longo da escolaridade do aluno *(portfólio)* pode constituir, ainda, material relevante para as decisões escolares e profissionais a tomar pelos alunos e seus encarregados de educação ao longo da escolaridade.

Em quarto lugar, uma das razões para a popularidade dos CRT no ensino prende-se com a progressiva consciência dos educadores (e dos movimentos de reforma educativa) da importância do ensino individualizado. Tal prática exige um sistema de avaliação freqüente e centrada no processo e nas aquisições da aprendizagem. Situação idêntica é defensável na educação especial. Importa que a avaliação sirva mais de perto o processo de ensino-aprendizagem, permitindo (i) a monitorização dos alunos e dos professores quanto aos progressos escolares, (ii) a definição dos objetivos para uma determinada intervenção, (iii) o diagnóstico de deficiências na aprendizagem, (iv) a avaliação de programas de intervenção sócio-educativa ou (v) a certificação das aprendizagens conseguidas pelos alunos.

Em quinto lugar, os CRT parecem muito mais adequados aos objetivos, em geral, de toda a avaliação escolar. Os CRT estão construídos já com a hipótese subjacente do crescimento na experiência ou nos conhecimentos, e daí mais orientados para avaliação dos ganhos progressivos intra-individuais - objetivo edumétrico - do que para a avaliação das diferenças estáveis entre os indivíduos - objetivo psicométrico (Carver, 1974). Gronlund (1973, ps. 42-45) refere várias situações de tal utilização: (i) avaliação inicial ou de diagnóstico (avaliar pré-requesitos necessários para a aprendizagem, determinar o lugar de cada aluno numa seqüência de aprendizagens, determinar a porção de uma área de instrução que os alunos já dominam, fornecer uma base para avaliação da aprendizagem adquirida durante a instrução, e sugerir formas concretas de superação de deficiências diagnosticadas); (ii) avaliação formativa ou contínua (determinar que objetivos de uma unidade foram atingidos, detectar deficiências com base nos itens falhados, prescrever as medidas julgadas convenientes para superar as dificuldades de cada aluno, e usar a informação para a promoção da própria instrução em termos de métodos, estratégias ou seqüências de aprendizagens); e (iii) avaliação sumativa (geralmente confinada ao final de um programa de formação e tendo essencialmente em vista a certificação dos indivíduos com bases nos produtos finais atingidos).

O uso dos CRT na avaliação sumativa ou de certificação final dos alunos requer, na maioria das vezes, que o nível de mestria fixado seja desdobrado e diferenciado. Se para a avaliação contínua era suficiente a fixação de um critério único de mestria, os objetivos de seriação e de classificação dos alunos no final de um ano escolar ou nível de escolaridade aconselham o recurso a provas diferenciadoras das aprendizagens e das competências desenvolvidas. Nestes casos, os testes centrados em critério poderão fixar vários níveis de realização ou de mestria. Por exemplo, podemos apontar um nível A ou um desempenho igual ou superior

a 90% na prova, um nível B ou 80% e um nível C ou 70% de resolução da prova (estes níveis não contrariam forçosamente a existência prévia de um nível geral de mestria da aprendizagem e realização - sujeitos com e sem mestria - que, para este caso, se situaria numa resolução a 70% da prova). Convem reafirmar que, também aqui, os níveis são fixados previamente, por outras palavras não se aguarda a realização dos indivíduos para se fixarem os valores (por exemplo fazer oscilar os valores de modo a que determinada percentagem de sujeitos obtenham os vários níveis indicados). Por outro lado, o teste de certificação inclui geralmente um maior número de objetivos a avaliar, um maior número de itens e um padrão de realização por norma mais exigente (Hambleton, 1982, ps.352-353). Em tais testes, o número de objetivos a avaliar é flexível, não é forçoso um número idêntico de itens para cada objetivo (sub-testes), o padrão de realização para fixação de mestria pode variar nos vários objetivos, a classificação final pode considerar a mera soma ou uma nota ponderada dos resultados nos sub-testes.

A popularidade dos CRT nos contextos de aprendizagem escolar conduziu à sua progressiva entrada nas empresas, mais concretamente no campo da formação profissional. A progressiva sistematização da formação profissional (aprendizagem programada, avaliação das competências mínimas, certificação profissional) e da avaliação de desempenho nas práticas de gestão dos recursos humanos e do desenvolvimento organizacional explicam o sucesso deste tipo de provas também na Psicologia das Organizações. Um dos exemplos dessa sua utilização prende-se com a construção de "amostras de trabalho" susceptíveis de avaliarem a formação e o desempenho do sujeito numa determinada área ou competência.

A utilização dos CRT nos contextos de formação, em geral, beneficia de algumas características destes testes e, sobretudo, do sentido prático da informação obtida para monitorar o próprio ensino e treinamento. A inserção destes testes nos objetivos da aprendizagem escolar ou da formação profissional parece-nos óbvia. Eles não só se adaptam, como decorrem dos planos curriculares (objetivos, conteúdos, estratégias). Ao mesmo tempo, o desempenho de cada sujeito é apreciado por referência ao nível de sucesso fixado e não fica condicionado aos níveis de desempenho dos pares (desde logo independentemente do aluno ou formando se encontrar integrado numa turma boa ou fraca). Os "norm-referenced tests" aparecem como menos informativos quando se pretende saber o que os sujeitos podem e não podem fazer, e fortemente condicionados na sua interpretação pela distribuição dos resultados no grupo em apreço (Glaser, 1963; Hambleton & Novick, 1973; Popham & Husek, 1969).

A este propósito podemos atender no significado diferente que as percentagens podem ter quando usadas num e noutro grupo de testes. Nos CRT, os resultados dos sujeitos são geralmente traduzidos em termos de percentagem de realização (por exemplo, o aluno consegue realizar 90% das tarefas apresentadas), o que se diferencia do uso da percentagem nos NRT (os percentis indicam quantos sujeitos se situam abaixo de determinado resultado).

3 - Construção dos CRT

Muito embora os CRT sejam hoje usados nas escolas e nas empresas, incidiremos a nossa análise no primeiro desses dois contextos. A maioria da biblio-

grafia disponível toma idêntica atitude, parecendo-nos ser um contexto mais universal a todos os leitores e, também, de mais fácil exemplificação ao longo do texto.

Alguns princípios gerais podem apontar-se à construção dos CRT (Gronlund, 1973, ps. 3-6), os quais terão que ser forçosamente ponderados na elaboração e seleção dos itens. Em primeiro lugar é necessário definir o domínio de uma forma clara e determinar as tarefas de aprendizagem. Por outras palavras, importa precisar o resultado esperado de uma aprendizagem a ser avaliado (âmbito). Esta definição pode ser mais facilmente conseguida em certas situações (por exemplo na aprendizagem da "aritmética") e mais difícil noutras (por exemplo na aprendizagem em "estudos sociais"). A construção de um teste avaliativo remete, de imediato, para a questão de "avaliar o quê?". A delimitação do âmbito do teste é melhor conseguida através da divisão do programa em unidades, e destas em sub-unidades, bem como da explicitação do tipo de comportamentos esperados (será diferente "adquirir conceitos" ou "aplicar princípios").

Algumas taxonomias existem, seja no sentido de orientar o ensino, seja no sentido de orientar a avaliação. Os testes educacionais centrados na avaliação do rendimento e da aprendizagem recorrem a tais taxonomias; veja-se a grande popularidade da taxonomia de Bloom entre os professores (Bloom *et al.*, 1956; veja capítulo 7 deste livro). É necessário ainda que as várias sub-unidades estejam corretamente representadas de modo a que os resultados dos sujeitos possam ser referenciados às mesmas.

Uma segunda exigência é a de que os objetivos educacionais devem ser claramente definidos em termos comportamentais ou de realização. Por exemplo deve atender-se às próprias condições em que determinado comportamento ou conhecimento deve ser demonstrado, servindo este esforço não apenas à avaliação dos efeitos do ensino-aprendizagem mas à sua própria programação. Mais uma vez, os itens devem cobrir objetivos esperados com a instrução, ou seja o que é esperado após uma determinada aprendizagem. Sendo os resultados nos testes interpretados em termos da realização numa determinada área, é necessário que os itens sejam medidas diretas dos resultados esperados da aprendizagem. Medidas indiretas da aprendizagem não são apropriados para os CRT.

Um terceiro aspecto tem a ver com a representatividade dos itens do teste, o que desde logo implica uma amostra de itens para avaliar um determinado domínio. Os itens reunidos devem refletir, de um modo adequado, os objetivos prosseguidos com o ensino. Mais à frente, a propósito da validade de conteúdo dos CRT, especificaremos melhor esta exigência.

Um quarto aspecto requerido prende-se com a definição clara dos padrões de realização tidos como desejáveis. Aqui as dificuldades prendem-se com as próprias características dos professores, exigências dos sistemas educativos nacionais ou das capacidades dos alunos em causa. Nalgumas situações tal padrão poderá corresponder a um tempo mínimo para a realização de certa tarefa (por exemplo, concluir determinadas operações em 15 minutos), à precisão da própria execução (por exemplo, realizar uma tarefa sem erros ou demonstrar firmeza no traço) ou ao número de respostas corretas (geralmente este último é o mais freqüentemente usado e corresponde à percentagem de itens corretos).

Os padrões de realização remetem-nos para a definição do nível de mestria. No final o teste deve permitir separar os sujeitos com mestria no comporta-

mento avaliado dos sujeitos que não atingem essa mestria. Este ponto não é estranho à generalidade dos professores quando elaboram os seus testes e os pontuam definindo os alunos com e sem aproveitamento no teste. No entanto, se o professor for confrontado com o pedido de explicitação da "competência mínima" ou competência exigida para a transição e sucesso na aprendizagem de um módulo seguinte, certamente que algumas dificuldades advêm. Este é um dos objetivos com a construção e uso dos CRT, muito embora exija um trabalho aturado na sua construção. O problema complica-se, ainda, quando a prova abarca diferentes subtarefas, variação de resultados em si mesmos (aquisição de conceitos, feitura de aplicações, desenvolvimento de características pessoais,...) ou quando se questiona a mestria não ao nível do aluno mas da turma.

3.1 - Construção, análise e seleção dos itens

Os itens do teste são mais analisados em relação à sua adequação e representação de determinado objetivo do que quanto ao seu índice de dificuldade ou poder discriminativo, como acontece nos testes formais ou centrados em normas. A pouca ênfase na diferenciação dos sujeitos entre si e nas normas de grupo para comparação dos resultados conduz a que esses dois parâmetros estatísticos de análise dos itens, essenciais à dispersão ou variância dos resultados, não sejam agora tão valorizados.

A qualidade de um item nos CRT é avaliada pelo grau em que reflete, em termos de conteúdo, os domínios dos quais foram retirados (Hambleton, 1982, 359). Para tal apreciação a opinião dos especialistas mostra-se essencial. Eles podem avalizar se a amostra de itens que compõe o teste cobre a listagem de comportamentos que se pretendem avaliar (ver "tabela de especificação" quando se falar na validade de conteúdo destes testes). Um outro procedimento freqüente para apreciar a qualidade dos itens consiste na administração prévia de tais itens junto de sujeitos similares àqueles a que o teste se destina e na análise dos resultados obtidos, nomeadamente em termos de proporção de respostas corretas. Adivinha-se, com estes passos e procedimentos de análise, que a construção dos CRT é morosa. A pouca ênfase colocada na inferência interpretativa dos resultados, como nos NRT, ou a possibilidade do seu uso mais generalizado em termos de sala de aula poderiam fazer supor, erradamente, poucas exigências ao nível da construção (Shaycoft, 1979).

Como o objetivo da avaliação pelos CRT é a obtenção de medidas de desempenho que possam ser expressas diretamente em termos de aprendizagem do aluno nas tarefas escolares, vários passos e cuidados devem ser seguidos e acautelados por parte dos seus construtores:

1) Definição e delimitação clara do *domínio das tarefas* de aprendizagem. Esta exigência é mais facilmente conseguida na instrução programada por pequenas unidades, em áreas como a aritmética (pior em áreas menos estruturadas como as línguas) e quando as competências de aprendizagem aparecem definidas para cada unidade (por exemplo, conhecimento de termos ou riqueza de vocabulário numa unidade de linguagem expressiva na língua estrangeira). Noutras situações o professor pode formar unidades tomando assuntos ou períodos de tempo (por exemplo unidades de uma ou duas semanas - Block, 1971). A avaliação de cada

unidade exige avaliações mais freqüentes, o que não deixa de trazer algumas vantagens ao processo de ensino-aprendizagem: (i) informação do aluno e do professor quanto às áreas de aprendizagem a merecerem maior atenção, e (ii) tipos específicos de dificuldades de aprendizagem e possíveis soluções. Por outro lado, como para se assegurar a fidelidade do teste importa ter vários itens, a avaliação por unidades evitará testes demasiado longos, os quais se podem tornar pouco pedagógicos e contraproducentes face aos objetivos da avaliação.

2) Os *objetivos instrucionais* devem poder ser definidos em termos comportamentais, por outras palavras, devem poder ser avaliados através de desempenhos diretamente observados. Definido o âmbito, importa agora especificar o comportamento requerido na realização (por exemplo, identificar um problema, descrever uma variável, construir uma hipótese, generalizar um resultado...). Também se podem incluir aqui as condições em que deve ocorrer a demonstração de um desempenho (por exemplo, se oral ou escrita, com e sem consulta de apontamentos e materiais,...). Os objetivos podem ser definidos mais em termos cognitivos (sabe, conhece, define,...) ou mais em termos comportamentais (escreve, efetua cálculos, realiza uma experiência no laboratório,...). Se o teste tem objetivos de diagnóstico, então as tarefas devem ser mais específicas. Na tabela 8-2 ilustra-se um teste de diagnóstico para a adição envolvendo "transporte" com números inteiros.

Tabela 8-2. Itens de adição com "transporte" em números inteiros

Domínio: Efetuar adições com números inteiros envolvendo transporte. **Mestria:** A criança tem que realizar corretamente 2 dos 3 itens em cada um dos 6 tipos de somas (Séries A, B, C, D, e E).
A. soma dois números inteiros de um dígito cuja soma é superior a 10 (ex.: 8+5)
B. soma três números inteiros de um dígito cuja soma seja superior a 10 (ex.: 5+7+2)
C. soma dois números inteiros com dois dígitos com transporte (ex.: 64+57)
D. soma dois números inteiros com três dígitos e repetidos transportes (ex.: 598+463)
E. soma três números inteiros com dois dígitos e repetidos transportes (ex.: 25+54+36)
F. soma três números com diferentes dígitos e repetidos transportes (ex.: 598+446+21)

3) Definição clara dos *padrões da realização*. Trata-se de um aspecto fundamental neste tipo de testes. Nalgumas áreas de realização da vida real eles estão definidos (por exemplo, define-se como analfabeto funcional a pessoa que não consegue ler uma notícia no jornal ou não consegue escrever uma mensagem simples). Importa fixar um nível de mestria no teste.

4) Definir um padrão de mestria para todos os alunos nas aprendizagens escolares é difícil; será mais correto definir um mínimo de competência necessária para transitar nas aprendizagens. Podemos tentar nesta definição responder a duas questões: (i) *o que deve ser a mestria* numa situação particular de aprendizagem; e (ii) *o que pode ser a mestria* numa situação particular de aprendizagem. Para a 1ª questão podemos tentar saber (i) que conhecimentos e competências mínimas são exigidas para o módulo de aprendizagem seguinte na mesma área?, (ii) que competências básicas numa área são pré-requisitos para a aprendizagem noutras áreas (por exemplo, que competências de leitura ou de cálculo)?, (iii) que míni-

mo de competências é exigido para a realização segura de uma dada atividade específica (por exemplo, conduzir um carro com segurança na cidade)?, (iv) que conhecimento e competências são necessárias para a eficiência mínima num trabalho (por exemplo, atender os clientes pelo telefone)?, ou (v) que conhecimento e competências mínimas são necessárias para funcionar no dia-a-dia?

Estas e outras questões ajudam-nos a fixar os resultados esperados da aprendizagem num dado nível e num área particular de aprendizagem; por outras palavras, ter critérios para definir uma aprendizagem sucedida. Esta definição de padrões, como se pode depreender pelas questões anteriores, não é fácil em bom número das aprendizagens. A questão não está apenas na delimitação do seu âmbito e na sua operacionalização comportamental. Também se deve considerar se os sujeitos podem ou não atingir esse nível de mestria (idade, proveniência, pré-requisitos,...). No fundo, importa assegurar um trabalho conjunto de professores, monitores, autoridades, especialistas de aprendizagem ou especialistas no curriculum, entre outros.

A fixação dos padrões de realização pode ser feita fixando um tempo limite para o desempenho, por exemplo ser capaz de visualizar as diferenças entre duas figuras em 3 minutos. Na sala de aula o critério tempo não é o mais usado. Usualmente o padrão é expresso em termos da percentagem de comportamentos a apresentar ou de itens do teste que se espera que os alunos respondam corretamente (Block, 1971, sugere que 80 ou 85% seria um padrão correto na fixação da mestria). Um nível mais elevado poderá desmotivar os alunos. Evidentemente que o valor da percentagem fixado depende do número de comportamentos avaliados. Com apenas três itens ou observações por situação, o nível de exigência na fixação de mestria pode situar-se em duas respostas corretas.

A fixação do padrão de desempenho não pode ser entendida como tarefa de um dado momento. Em termos escolares, por exemplo, esse padrão deverá ser fixado e analisado anualmente. Na falta de outros elementos, poder-se-á atender: (i) à dificuldade dos itens incluídos (esperar: 80% de correto nas questões curtas; 85% nas de escolha múltipla; e 90% nas questões "verdadeiro-falso" - esta percentagem diferenciada procura atender à possibilidade de respostas certas por acaso); (ii) se um nível superior de mestria é necessário para o sucesso na fase de instrução seguinte ou por razões de segurança no uso de equipamento do laboratório, então deve subir-se a percentagem exigida de respostas corretas; (iii) se as unidades seguintes vão repetir conhecimentos e competências de uma unidade anterior, então a avaliação desta não requer um padrão tão elevado de respostas corretas.

5) A realização do aluno deve constituir uma *amostra adequada de comportamentos* dentro de cada área de realização. Um teste é sempre uma amostra de tarefas selecionados para representar um domínio mais largo do comportamento. Em aprendizagens muito simples (conhecer as vogais, contar até 10,...) é possível definir e avaliar o comportamento no seu todo. Em áreas mais complexas, ou mais abrangentes, apenas uma amostra limitada de realizações pode ser tomada na avaliação. Nestes casos, a melhor representatividade de tal amostra é conseguida quando (i) a aprendizagem é dividida em unidades relativamente pequenas, (ii) o âmbito das tarefas de aprendizagem é claramente definido, (iii) os passos específicos são tomados para se obter uma amostra adequada. O recurso a "tabelas de especificação" é o meio mais adequado (quadro de dupla entrada

fixando os objetivos instrucionais e as áreas de conteúdo), devendo o número de itens em cada cela da tabela refletir a ênfase instrucional da unidade - claro está que, quantos mais itens tivermos por cada especificação, mais influência vão ter no resultado final e, também, mais possibilidades de informação teremos.

6) Os itens são selecionados consoante o grau em que refletem bem os *comportamentos especificados* nos objetivos instrucionais. Como os CRT são interpretados em termos de padrões absolutos de realização num dado conjunto de tarefas de aprendizagem, então é importante que os itens do teste sejam uma medida direta dos resultados esperados da aprendizagem. O comportamento representado por cada item deve cobrir um comportamento específico descrito nos objetivos instrucionais. Quando esse esforço existe, então podemos generalizar a partir do teste para o mais largo domínio de tarefas de aprendizagem que o teste representa. Medidas indiretas da aprendizagem (veja-se a componente inferência na interpretação dos resultados nos NRT) ou selecionar os itens com base no índice de dificuldade são inapropriados. A dificuldade nos CRT deve apenas decorrer da natureza das tarefas de aprendizagem a avaliar e não de qualquer tentativa de se obterem escores diferenciados como nos NRT (leque variado de índices de dificuldade dos itens para melhor diferenciação dos sujeitos entre si). De fato, quando a mestria é o objetivo da formação, então todos ou quase todos os sujeitos devem responder corretamente aos itens. Nos CRT os itens não deveriam ser resolvidos por qualquer sujeito no início da instrução e por todos ou quase todos após a instrução (quer o item quer a instrução foram efetivos). A variação dos resultados nos CRT não é relevante. Os itens não têm que ser fáceis nem difíceis; devem adequar-se aos critérios instrucionais definidos.

7) Um *sistema de cotação* que descreva de forma adequada o nível de eficiência do aluno nas tarefas de aprendizagem. Os CRT pretendem descrever o desempenho do aluno em termos absolutos, ou seja, a informação é já em si mesma significativa mesmo sem qualquer referência ao desempenho dos seus pares. O escore deve indicar, precisamente, que tarefas de aprendizagem o aluno pode realizar num dado nível de eficiência, daí que as notas percentílicas ou os resultados padronizados, como nos NRT, são inapropriados. Estes indicam a posição relativa do sujeito no seu grupo de pares, mas não define o nível de eficiência em que o sujeito se encontra. Nos CRT o nível de eficiência é calculado independentemente da realização dos demais membros do grupo, mais ainda não é influenciado pelo fato do grupo ser maioritariamente constituído por sujeitos bons ou fracos realizadores. É diferente dizer que o aluno pode definir corretamente 90% dos termos incluídos na unidade (o resultado em "percentagem correta" é muito usado nos CRT) ou dizer que ele se situa no percentil 90. Neste último caso, a apreciação não pode ser feita sem considerar o leque de distribuição das notas e as posições dos outros sujeitos.

Finalmente alguns cuidados pontuais devem ser considerados no momento da elaboração da versão final de um teste centrado em critério. Tais cuidados retomam os pontos abordados antes, sobretudo se esta revisão final é feita por alguém que não participou nas fases anteriores de construção da prova (análise e revisão final do teste que se aconselha). Entre outros autores, Gronlund (1973) sugere alguns cuidados:

a) revisão da relevância dos itens para os resultados da aprendizagem (ver se os itens correspondem a comportamentos específicos por sua vez traduzindo resultados da aprendizagem a avaliar);
b) revisão da redação dos itens (por exemplo, as suas ambigüidades, o tipo de terminologia ou a existência de informação não relevante para o problema em causa, frases longas que requerem muita leitura e interpretação não sendo isso o objetivo);
c) atenção a fatores irrelevantes na formulação das questões e a pormenores nas alternativas de resposta que possam induzir a resposta nesse item ou em itens posteriores, ou que impliquem dificuldades acrescidas;
d) verificar se, no seu conjunto, os itens correspondem à "tabela de especificações" elaborada, e se o seu número é tido como suficiente para a amostra de comportamentos que se pretende avaliar;
e) agrupar o conjunto de itens que avaliam um mesmo objetivo, e organizar internamente os itens dos mais fáceis para os mais difíceis, seja dentro de cada unidade, seja do conjunto do teste;
f) colocar os itens na folha da prova de modo a uma mais fácil resposta por parte dos sujeitos e à maior facilidade de correção dos resultados, e enumerá-los consecutivamente;
g) escrever instruções claras para o teste no seu todo e para cada unidade, quando necessário, fazendo com que a tarefa do sujeito a realizar fique clara e definida;
h) redigir o item de forma positiva sempre que possível (assentar nos fatos e não nas excepções, o aluno pode saber que não é o caso mas nada nos assegura que saiba o que é o caso, alguns alunos respondem aos itens não lendo o "não");
i) cuidar de aspectos específicos que podem induzir a resposta (concordâncias de gênero e número, associações verbais que tornam a resposta correta óbvia, os advérbios de tempo "algumas vezes" ou "sempre" que aumentam a probabilidade de uma afirmação estar certa ou errada, respectivamente, ou inclusão de miscelânea de fatores,...);
j) confirmar se os elementos do júri ou corpo de especialistas dão o mesmo sentido aos itens e se são unânimes na indicação da resposta correta.

Usualmente dois tipos básicos de itens surgem nos CRT: os itens de escolha e os itens de produção de resposta. Neste último caso incluem-se itens de completamento (sem indicação de alternativas), respostas breves, respostas de desenvolvimento e composições. No primeiro caso, podemos falar nos itens de emparelhamento, verdadeiro-falso, e escolha múltipla. Estes formatos devem estar também de acordo com o tipo de resultado da aprendizagem que se quer avaliar. A resposta curta é aconselhável para definições, nomeações ou descrições de algo. Os itens de emparelhamento podem ser úteis quando se trata de identificar ou de distinguir entre objetos e conceitos (no entanto este formato exige séries de coisas homogêneas, por exemplo: datas - eventos, autores - livros, instrumentos - usos, países - capitais, etc.). A composição pode ser aconselhada nos itens que pretendam avaliar, por exemplo, os estilos de comunicação ou a criatividade dos sujeitos.

3.2 - Características metrológicas dos resultados nos CRT

Os métodos tradicionais de cálculo da sensibilidade, fidelidade e validade não têm aplicação dada a pouca ênfase dos CRT na diferenciação dos sujeitos entre si. Em consonância, a variância dos escores encontra-se diminuída. Por sua vez pode não ser defensável uma grande homogeneidade dos itens entre si. As situações de aprendizagem podem apresentar-se diversificadas, e logicamente também os itens propostos para a sua avaliação.

A apreciação das características metrológicas dos CRT deverá ser feita mais pela apreciação dos itens e dos dados do que pelo cálculo de coeficientes estatísticos. Um certo esforço deve ser feito em ordem à melhoria das características metrológicas destes instrumentos. Fatores vários afetam tais coeficientes e é necessário atender à sua existência: formato dos itens, a sintaxe e a semântica dos termos ou idéias expressas, as variáveis de situação ao nível dos sujeitos (fadiga, interesse, maior aprendizagem em certos contextos), das condições de espaço e de tempo em que decorre a avaliação ou as condições do próprio observador (Nitko, 1980a, ps. 54-55).

Em relação à sensibilidade, podemos afirmar que os CRT não se orientam para a obtenção de uma distribuição dos resultados de desempenho de acordo com as leis da "curva normal". No limite, os CRT permitem apenas separar dois grupos, mesmo assim com um número de efetivos bastante desigual. Referimo-nos à separação única entre o grupo dos alunos com mestria (a grande maioria) e o grupo dos alunos sem mestria.

Ao nível da fidelidade dos resultados, várias dificuldades se colocam ao recurso da metodologia consagrada nos testes formais ou referenciados em normas. Tais métodos, todos de índole correlacional, estão influenciados pela variância que se encontre nos resultados. Essa variância nos CRT é reduzida dada a pouca dispersão dos resultados. A não inclusão de itens com índices de dificuldade variados, o fato dos CRT serem geralmente curtos no seu tamanho (um só item poderá ser suficiente para a avaliação de determinado objetivo), a justa expectativa de que nas situações normais de ensino a generalidade dos alunos falhem o teste antes da aprendizagem de determinado domínio e acertem depois dessa aprendizagem efetuada, por exemplo, justificam a pouca dispersão dos escores. A inserção dos CRT no processo de ensino-aprendizagem justifica que na seleção dos itens se procure sobretudo atender ao seu grau de sensibilidade para com o progresso na própria aprendizagem e ao seu grau informativo para a prossecução do ensino-aprendizagem junto de cada indivíduo.

O uso do modelo teste-reteste no cálculo da fidelidade dos resultados pode ser adequado nos CRT. Esse cálculo não será feito através do coeficiente de correlação produto-momento como nos NRT, mas do somatório das proporções dos sujeitos que no teste e no reteste são classificados com mestria e com não mestria. Fidelidade traduz aqui a proporção de acordo ou de examinandos consistentemente classificados (com mestria e sem mestria) nas duas aplicações do teste (Hambleton, 1982, p. 364; Hambleton & Novick, 1973). Idêntica fórmula pode ser aplicada quando a avaliação é feita através de dois observadores independentes (aqui o acordo é feito com base nos resultados dos dois observadores e não dos dois momentos de

avaliação). Se em testes mais orientados para a individualização do ensino podemos aceitar proporções de acordo em torno de 75%, naqueles que visam certificar os alunos (diplomas) ou habilitar profissionalmente um indivíduo (licença para uma determinada prática profissional) essa margem de acordo deve aproximar-se de 90%. Na tabela seguinte ilustramos o cálculo dessa margem de acordo (p_o):

		2ª aplicação	
		Mestria (i)	Não Mestria (k)
1ª aplicação	Mestria (i)	ii	ik
	Não Mestria (k)	ki	kk

$p_o = S\ p_{ii} + p_{kk}$ (proporção de sujeitos simultaneamente classificados com mestria e não mestria)

Segundo Swaminathan, Hambleton e Algina (1974), a fórmula apresentada tem a desvantagem de não considerar a proporção de acordo fruto do mero acaso. Estes autores sugerem uma fórmula corretiva (uso do coeficiente k de Cohen, 1960), mais concretamente $k = (p_o - p_c) / (1 - p_c)$. Neste caso, p_c traduz a proporção de acordo que poderia ocorrer se as classificações nas duas aplicações seguisse uma ordem estritamente aleatória.

Tomando a tabela anterior e supondo os valores seguintes na distribuição dos sujeitos pela mestria e não mestria nas duas aplicações do teste,

		2ª aplicação		
		Mestria	Não Mestria	Total
1ª aplicação	Mestria	18	2	20
	Não Mestria	2	8	10
	Total	20	10	30

teríamos,

$p_o = 18/30 + 8/30 = .866$
$p_c = 20/30 \,.\, 20/30 + 10/30 \,.\, 10/30 = .555$
$k = (.866 - .555) / (1 - .555) = .70$

Como nos NRT, a fidelidade dos resultados aparece afetada pelo número de itens ou tamanho do teste, pela qualidade dos próprios itens ou pela heterogeneidade dos grupos tomados nesse cálculo. Alguns destes aspectos não são devidamente contemplados nos CRT. Geralmente estas provas têm um número mais reduzido de itens e, mesmo que os grupos possam ser originariamente heterogêneos, eles ficam apenas diferenciados em "mestria" e "não mestria". Por último, importa atender ao nível de exigência em que foi colocada, para um dado teste específico, a nota padrão que fixa a mestria.

Nos testes em que a mestria está diferenciada por três ou quatro níveis (situação mais freqüente nos testes destinados a avaliações sumativas ou de certificação/seleção dos alunos) poder-se-á recorrer igualmente à metodologia do teste-

reteste. Neste último caso podemos recorrer ao coeficiente de correlação para variáveis ordinais, ou seja o *Rhô de Spearman (rank-order correlation).*

Em relação à análise da validade dos resultados nos CRT importa mencionar o grande recurso que é feito à validade de conteúdo (Gronlund, 1973, p. 47), até porque se exige serem estes testes medidas de áreas de aprendizagem específicas. A construção de um CRT requer a definição clara do domínio de conhecimento ou das competências a serem avaliadas e dos objetivos do teste. Essa definição justifica a necessidade de previamente se elaborar uma "tabela de especificações" orientadora do trabalho de construção, análise e seleção dos itens.

A "tabela de especificações" deve integrar, para além dos conteúdos das aprendizagens efetuadas, os objetivos prosseguidos numa unidade educativa: compreensão de termos, compreensão de fatos, compreensão de conceitos, compreensão de princípios ou a compreensão de procedimentos, ou então centrar-se não na compreensão mas na aplicação ou na interpretação. Os itens recolhidos devem merecer o acordo de especialistas nesse âmbito quanto à sua representatividade em relação a ambos os aspectos (conteúdos e objetivos).

Nestes testes fazem sentido, ainda, estudos de validação no sentido da validade de construto. A validade de conteúdo apenas nos assegura a representatividade dos itens em relação ao domínio ou aos objetivos que se pretende avaliar. A validade de construto é essencial para estabelecer a validade das descrições e decisões feitas com base nos resultados em testes centrados em critério (Hambleton, 1982, 371). Por exemplo, poder-se-á questionar se os itens apenas avaliam a informação possuída pelo aluno ou se entram nas áreas de compreensão de conceitos. Dado que um e outro aspecto têm significações diferentes é necessário salvaguardar a "qualidade" ou a significação da informação que os resultados podem traduzir. Aliás, ao mesmo tempo que as inferências pedagógicas se baseiam nos resultados obtidos pelos sujeitos, certo também que os resultados são uma função das respostas dos sujeitos (Messick, 1975, ps. 960-961). Aqui pode, inclusive, questionar-se o critério utilizado para se ter mestria em determinado objetivo, pelo menos discutir a sua adequabilidade e justificação.

Finalmente, é necessária alguma informação sobre a relação entre os resultados no teste e outras situações de realização dos sujeitos (validade referenciada em critérios externos). Aspecto essencial na utilização prática dos CRT em educação é o da possibilidade de previsão do sucesso posterior do sujeito, por exemplo nas próximas unidades do programa ou objetivos, a partir da sua realização no teste. Análises correlacionais, ou o recurso aos grupos diferenciados (com e sem mestria), podem utilizar-se para estas análises. Os resultados agora obtidos poderão também dar alguma informação a propósito da validade de conteúdo e de construto do teste.

Ao nível das normas de interpretação dos resultados, enquanto os NRT enfatizam o seu cálculo com base nos parâmetros de realização do grupo (média, desvio-padrão). e assim aparecem os percentis, as notas Z, as notas T e as classes normalizadas de distribuição dos resultados, nos CRT temos essencialmente as percentagens de sucesso, as categorias e os resultados em termos de mestria ou não-mestria (fixação de pontos de corte). Não existe a necessidade de uma referência estatística ou a comparação com um grupo normativo. O resultado no CRT tem um significado em relação a um critério objetivo e que é independente das realizações

individuais. Importa, aqui, cuidar da qualidade e significado como o ponto de mestria singular ou plural (vários índices de mestria) foi calculado. Vários métodos servem este objetivo, desde o simples julgamento por especialistas até à tomada de uma nota que permita a separação de dois grupos contrastantes no domínio a avaliar, passando pelas fixações dos pontos de corte *(cut-off escore)* em função dos propósitos da própria avaliação.

4 - Dificuldades dos CRT

Mencionaremos algumas dificuldades dos CRT na avaliação psicopedagógica, partindo de problemas que lhes são inerentes e de exigências práticas que lhes são externas (Ebel, 1972). Em primeiro lugar, é necessário não confundir as diversas expressões associadas à designação destes testes (Nitko, 1980b). Por exemplo confunde-se muitas vezes o termo "critério" com o "resultado de passagem", o nível de realização previamente fixado para atribuição de mestria *(cut-off escore)* ou o critério de exigência (Hambleton, 1982). O mais importante nos CRT é o domínio em avaliação (sem qualquer referência ao domínio ou aos objetivos em avaliação, nada pode ser afirmado a propósito da realização dos indivíduos nesse domínio). Assim, o termo "critério" foi desde o início (Glaser, 1963) usado para descrever um domínio de conteúdo ou de comportamentos aos quais se podiam referenciar os resultados nos testes (Hambleton, 1982). No entanto, associar "critério" a "nível de mestria" parece ser mais fácil do que a "conteúdo", daí a confusão freqüente. Anastasi (1982, 95) pensa, por exemplo, que a designação mais adequada para os CRT seria "testes centrados no conteúdo".

Ligado a este aspecto, temos a expressão *"mastery test"* que alguns autores identificam com os CRT. Tal identificação pode considerar-se abusiva pois apenas considera a exigência técnica da realização do sujeito ser classificada em termos de atingir ou não atingir a "mestria". A ênfase é colocada no nível de realização e não no domínio, fazendo perder todo o sentido da própria avaliação. Apenas nos interessa saber se o sujeito tem mestria se identificamos bem em quê ou, por outras palavras, não é possível pensarmos que o sujeito tem mestria sem uma referência a um determinado domínio. Além disso, a expressão *"mastery test"* pode ter subjacente um processo de ensino-aprendizagem seqüencial específico *(mastery learning)*, que nem sempre acontece quando os CRT são utilizados.

Uma segunda expressão a diferenciar é a de *"criterion-related validity"*. Trata-se de uma metodologia ou orientação na análise da validade dos resultados num teste (validade externa) e não tem a ver com a caracterização dos CRT. Tal metodologia aparece utilizada também, aliás sobretudo, nos NRT e procura informar-nos da capacidade preditiva que tais resultados apresentam em relação às realizações dos sujeitos noutras situações. Também nos CRT tal metodologia aparece utilizada, designadamente quando o objetivo é selecionar os sujeitos para diferentes programas ou alternativas escolares (apesar da grande utilização dos CRT ser movida por preocupações "aqui e agora" e não tanto para a tomada de decisões em relação ao futuro dos indivíduos).

Uma terceira expressão freqüentemente usada é a de *"domain-referenced testing"* (Millman, 1974). A diferença com os CRT é mínima, apesar da

generalidade dos autores optarem por esta última (Glaser, 1963; Hambleton, 1982, p. 352). Também em sentido próximo aparece a designação *"objective-referenced testing"*, que sendo uma expressão mais específica que *"criterion-referenced testing"* salienta a objetividade necessária na formulação dos itens (os itens como objetivos comportamentais).

A par destas dificuldades, mais em termos de terminologia ou dos conceitos, outras podem apontar-se em relação à metodologia de avaliação referenciada em critério. Mencionaremos algumas necessidades da prática a que os CRT poderão ter algumas dificuldades de resposta. Em primeiro lugar certas situações da prática psicopedagógica requerem alguma informação sobre a probabilidade de êxito futuro dos sujeitos e, por vezes, alguma comparabilidade dos desempenhos inter-sujeitos. Ora, não tendo a generalidade dos CRT sido construídos com tais preocupações, dificilmente eles podem responder cabalmente nessas situações. É possível, aliás, que existam alguns problemas de extrapolação do uso destes testes para certas variáveis psicológicas e situações da prática (domínios psicológicos e variáveis ainda não completamente descritos, por exemplo). Estes pontos fazem-nos pensar na complementaridade que os CRT e os NRT podem desempenhar na avaliação, até porque um e outro podem aproximar-se reciprocamente ao nível da interpretação que é feita dos resultados (Hambleton *et al.*, 1978).

Finalmente, existem algumas dificuldade em fixar o critério de mestria a atingir (divergência entre os autores ou problema jamais resolvido em definitivo - Gronlund, 1973, p. 13). Não é fácil fixar um critério que possa servir a todos os alunos de estímulo para a aprendizagem. Para alguns ele pode ser pouco desafiador e para outros ser ocasião de uma derrota pessoal constante. Aliás, neste último caso, nem sempre a não satisfação do critério significa a impossibilidade de prosseguir num programa ou que a mera repetição das tarefas possa levar à sua resolução.

Block (1971) defende para a generalidade das situações escolares um padrão de realização entre 80% e 85%, considerando este padrão mais realista. Em sua opinião, a proposta de um critério mais elevado pode afetar a motivação dos alunos. Gronlund (1973, p. 12) acrescenta que o valor numérico de tal padrão não pode ser separado do formato dos próprios itens formulados e sugere que nas situações de respostas a elaborar pelo sujeito (questões breves) esse padrão seja de 80%, sendo de 85% ou de 90% consoante o formato dos itens seja de "escolhas múltiplas" ou respostas "verdadeiro-falso", respectivamente (aspecto que nos remete para a probabilidade de respostas corretas devidas ao acaso).

Uma outra pista na determinação do nível de mestria consiste no recurso a algum critério externo disponível, melhor ainda se existir já uma forma paralela do teste. A observação direta do desempenho do sujeito pode definir, ainda, um desempenho aproximado, isto é o nível de tolerância sem afetar a atribuição da mestria no final. Isto é ainda mais objetivo se realizado através do recurso a vários avaliadores independentes. Por último, o nível de mestria pode ser fixado a título provisório pelo avaliador no momento da construção do teste, confirmando-o depois nas primeiras aplicações, por exemplo tomando o nível atingido pelos alunos no final da instrução daquela unidade curricular.

Claro está que a questão da fixação do critério não fica cabalmente resolvida. Os pontos anteriores são mais claramente aplicados em aprendizagens res-

tritas ou em unidades curriculares bem delimitadas. Se dos objetivos exclusivamente centrados na aprendizagem (conhecimentos, competências,...) passarmos aos objetivos de desenvolvimento psicossocial da Educação e da Escola, a tarefa complica-se. Aliás, um problema similar se pode encontrar quando as aprendizagens se reportam a conhecimentos menos básicos, por exemplo aquisições em termos do pensamento crítico, da criatividade ou da capacidade argumentativa dos alunos. Nestas situações pode ser mais difícil e menos interessante fixar a "mestria". Possivelmente, havendo NRT disponíveis para a avaliação de tais construtos, eles poderão melhor satisfazer as necessidades de avaliação. As pontuações de um grupo de alunos num teste formal de criatividade podem informar-nos sobre a eficácia de um programa de treino da criatividade.

5 - Conclusão

A partir dos anos 60 surge na Psicologia, nomeadamente na Psicologia Educacional, um novo tipo de testes de avaliação, designados *"criterion-referenced tests"* (Glaser, 1963). Este grupo de testes constitui-se em alternativa aos *"norm-referenced tests"*. Uma principal diferença se estabelece entre estes dois grupos de testes. Os testes centrados em normas fundamentam a interpretação dos resultados individuais nos valores da média e desvio-padrão na população, os testes centrados em critério informam da realização do indivíduo sem referência à realização dos outros, antes classificam os indivíduos como tendo atingido ou não determinado padrão de realização previamente definido num domínio específico (Black, 1985). O resultado é analisado no quadro do desempenho numa dada competência *(skill)*, bem definida, mais do que através da comparação do nível atingido por comparação com o nível dos outros sujeitos. Por estas razões, os testes referenciados a critério indiciam para uma avaliação menos comparativa e competitiva entre os sujeitos avaliados. Na tabela 8-3 descrevem-se sumariamente as características dos dois grupos de testes.

Ao longo deste capítulo fomos fazendo referência aos elementos que melhor descrevem os CRT e como tais elementos acabam por justificar uma maior utilização deste grupo de testes na avaliação da aprendizagem, da formação e do desempenho. Algumas diferenças foram sendo apontadas tomando estes testes comparativamente aos testes mais clássicos na Psicologia (os testes referenciados a normas). Podemos afirmar existirem duas formas habituais de ponderar o resultados individuais em situações de desempenho: (i) como realiza ou em que lugar fica comparativamente aos seus pares; (ii) quanto manifesta possuir no *construto* que está a ser avaliado. Este segundo aspecto, em que assentam os CRT, é mais difícil de se conseguir em áreas tradicionais da avaliação, nomeadamente quando a definição dos construtos internos não se encontra minimamente estabelecida e onde, mesmo com alguns avanços metodológicos, não podemos fixar unidades-padrão de medida. A opção nos testes clássicos da Psicologia foi, e continuará a ser, a ponderação assente na comparabilidade interindividual dos desempenhos obtidos em amostras representativas da população.

A lógica comparativa interindivíduos pode merecer alguns reparos numa sociedade mais liberal e democrática. Uma alternativa plausível de interpretação dos escores individuais é, precisamente, em termos do tipo de comportamento ou reali-

Tabela 8-3. Elementos comparativos dos testes referenciados a normas (NRT) e a critério (CRT)

Categoria	Teste referenciado a normas	Teste referenciado a critério
Planeamento do teste	Definição de um construto, por norma dimensão interna, genérica e abrangente (traço)	Definição dos objetivos da aprendizagem ou da realização num domínio de competências bem delimitado
Finalidade do teste	Determinar a posição do sujeito no seu grupo; seriar ou selecionar os sujeitos com base nos resultados; comparação interindividual dos desempenhos	Apreciar o que o sujeito conhece num determinado conteúdo; diagnosticar problemas específicos de aprendizagem; avaliar a eficácia de programas de treino ou ensino
Qualidade dos itens	Máxima discriminação entre os sujeitos; índice de dificuldade variável, sendo a maioria dos itens de dificuldade moderada (25 e 75% de respostas corretas) e correlação item-total acima de .30	Apreciar a eficiência dos sujeitos num dado domínio (amostra); Dificuldade variável mas a maioria dos sujeitos responde corretamente
Estandardização	Procura-se evitar a aprendizagem e o treino anterior; tentativa de avaliação da capacidade máxima (potencial); criação de uma situação laboratorial de avaliação	Avalia-se o treino e a aprendizagem anterior; centrados na realização habitual; avaliação nos contextos usuais de aprendizagem ou de realização
Parâmetros de apreciação	Apreciação considera a média e o desvio-padrão dos resultados que se obtenham num grupo de referência; proporção de sujeitos acima e abaixo de um indivíduo	Os parâmetros de apreciação são prévios e situam o sujeito em relação aos níveis especificados (critério de mestria); proporção de conteúdos que o sujeito aprendeu daquilo que era esperado
Validade dos resultados	Capacidade preditiva da informação recolhida para situações externas ao teste; itens mais genéricos para aumentar a sua capacidade de generalização	Resultados referenciados aos objetivos da aprendizagem; itens definem o que está ou não aprendido com fortes aplicações à individualização educativa

zação que um aluno é capaz de demonstrar ou apresentar (por exemplo, para ter a carta de condução não pode deixar ir o carro abaixo mais que duas vezes e não pode tocar no passeio mais que uma vez num estacionamento durante o exame; ou para ter o diploma de dactilógrafa deve conseguir escrever 50 palavras por minuto, etc.). Neste casos, como se depreende, a avaliação de cada indivíduo é feita mediante um padrão comum, ou seja um nível de realização tido como absoluto, independente das situações e das posições relativas dos demais sujeitos, sendo prévio à própria avaliação em que é usado (Glaser, 1971).

Estas características tornam o processo de construção e validação dos CRT moroso e com custos materiais e humanos significativos. Concordaremos que a avaliação dos CRT é "tendencialmente um processo nunca acabado" (Black & Dockrell, 1984, p. 66). Tomando procedimentos mais lógicos ou mais empíricos podemos apreciar, em primeiro lugar, a qualidade de tais instrumentos colocando algumas questões ao nível dos respectivos itens. Por exemplo, importa apreciar se os itens repre-

sentam um domínio definido previamente, se são sensíveis à instrução ou formação havida, se não incluem aspectos acessórios que possam confundir o seu significado e a significação do seu desempenho, se reportados a um mesmo critério apresentam entre si bom índice de homogeneidade.

Um segundo aspecto a considerar na avaliação dos CRT é a opinião dos próprios especialistas na área. Um bom CRT não pode prescindir da sua análise. Professores da mesma disciplina e curso, que não elaboraram os itens, podem apreciar a adequação da amostra de itens propostos para avaliar uma determinada competência ou aprendizagem. No final, podemos reter apenas os itens que reuniram o consenso dos elementos do júri.

O recurso a juizes oferece-nos outras vantagens para a validação dos CRT. A apreciação dos elementos do júri pode ainda incluir a especificação do tipo de comportamento avaliado com cada item (definição de conceito, diferenciação de conceitos, aplicação de conceitos, generalização de conceitos,...) e a estimativa do índice de dificuldade ou facilidade (% de alunos que não terão dificuldade em responder ao item). No final estes dois elementos informativos podem entrar igualmente na escolha dos itens a reter no teste.

As potencialidades práticas destes testes são evidentes, mesmo que não possam atender a todas as solicitações da prática. Dois pontos críticos tendem a ser mencionados a este propósito: (i) a pouca margem de predictibilidade que os resultados nestas provas tendem a expressar face a uma decisão de orientar a avaliação para as aquisições num determinado domínio e momento, e (ii) a reduzida variabilidade dos desempenhos individuais dado o fim último destas provas se confinar à classificação dos sujeitos no grupo de mestria e sem mestria.

A concluir, sendo o uso dos instrumentos de avaliação determinado pelas finalidades dessa mesma avaliação, a opção por CRT ou por NRT deve ser ponderada pelo profissional. Infelizmente poucos são os países que possuem os dois tipos de provas disponíveis. Julgamos que cada grupo de prova tem a sua especificidade na construção, mesmo que no limite possamos concordar com Hambleton quando afirma poder ser o formato idêntico nos dois tipos de testes, ambos poderem ser estandardizados e os seus resultados padronizados por referência a normas (Hambleton, 1982). Em nossa opinião faz pouco sentido tentarmos "adulterar" a respectiva filosofia para que um dos grupos possa apreender e fornecer parte da informação do outro. Consoante as decisões a tomar, o profissional da avaliação deve decidir do tipo de provas a usar e, inclusive, recorrer a ambas em muitos dos casos.

Bibliografia

Almeida, L. S., Ribeiro, I. S. & Correia, L. M. (1994). Testes centrados em critério: A sua incidência em educação. *Psicologia, IX* (3), 361-367.

Anastasi, A. (1982). *Psychological testing.* New York: Macmillan.

Black, H. D. & Dockrell, W. B. (1984). *Criterion-referenced assessment in the classroom.* Edinburgh: SCRE.

Black, H. D. (1985) Whither research on criterion-referenced assessment? *Research Intelligence, 19*, 2-5.

Block, J. H. (1971). *Mastery learning: Theory and practice.* New York: Holt, Rinehart and Winston.

Bloom, B. S. et al. (1956). *Taxonomy of educational objectives.* New York: Longman.

Carver, R. P. (1974). Two dimensions of tests: Psychometric and edumetric. *American Psychologist, 29*, 512-518.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Ebel, R. L. (1972). Some limitations of criterion-referenced measurement. In G. H. Bracht, D. K. Hopkins & J. C. Stanley (Eds.), *Perspectives in educational and psychological measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Glaser, R. & Nitko, A. J. (1971). Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational measurement.* Washington, DC: American Council of Education.

Glaser, R. (1963). Instructional technology and the measuring of learning outcomes: Some questions. *American Psychologist, 18*, 519-521.

Glaser, R. (1971). Instructional technology and the measurement of learning outcomes. In W. J. Popham (Ed.), *Criterion-referenced measurement.* Englewood Cliffs, NJ: Educational Technology Publications.

Gray, W. M. (1978). A comparison of piagetian theory and criterion-referenced testing. *Review of Educational Research, 48*, 223-249.

Gronlund, N. E. (1973). *Preparing criterion-referenced tests for classroom instruction.* New York: Macmillan.

Hambleton, R. K. & Novick, M.R. (1973). Toward an integration of the theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10*, 159-171.

Hambleton, R. K. (1982). Advances in criterion-referenced testing technology. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology.* New York: John Wiley and Sons.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 48*, 1-47.

Messick, S. A. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*, 955-966.

Millman, J. (1974). Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education: Current applications.* Berkeley, CA: McCutchan.

Nitko, A. J. (1980a). Criterion-referencing schemes. *New Directions for Testing and Measurement, 6*, 35-71.

Nitko, A. J. (1980b). Distinguishing the many varieties of criterion-referenced tests. *Review of Educational Research, 50*, 461-485.

Popham, W. J. & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*, 1-9.

Popham, W. J. (1971). Indices of adequacy for criterion-referenced tests items. In W. J. Popham (Ed.), *Criterion-referenced measurement.* Englewood Cliffs, NJ: Educational Technology Publications.

Popham, W. J. (1975). *Educational evaluation.* Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Shaycoft, M. F. (1979). *Handbook of criterion-referenced testing: Development, evaluation, and use.* New York: Garland STPM.

Swaminathan, H., Hambleton R. K. & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement, 11*, 262-267.

TESTS INFORMATIZADOS

Gerardo Prieto Adánez

1 - Introducción

El término *test informatizado* se emplea en la actualidad para describir un tipo de test en el que el ordenador es el soporte de todas las fases de la ejecución de la prueba:

(a) presentación en pantalla de las instrucciones
(b) presentación en pantalla de los ejemplos de práctica
(c) presentación en pantalla de los ítems
(d) registro de datos censales emitidos mediante el teclado o el ratón (nombre, edad, sexo, etc)
(e) registro de las respuestas emitidas mediante el teclado, el ratón, micrófono, o pantalla sensible al tacto
(f) codificación numérica de las respuestas
(g) almacenamiento de los datos
(h) puntuación de la prueba
(i) emisión de informes sobre los resultados.

Por tanto, no ha de aplicarse el calificativo de informatizado a los tests de papel y lapiz que se contestan en hojas de respuesta legibles mediante una lectora óptica conectada a un ordenador.

El origen y la proliferación de este tipo de tests es paralela a la difusión de los ordenadores personales. Desde la década de los 70 con las primeras versiones informatizadas del MMPI (Lushene, O'Neil, & Dunn, 1974), el crecimiento de esta metodología ha sido exponencial en la psicología y la educación. La publicación por la APA de una normativa para la administración e interpretación de los tests informatizados se ha debido a la importancia y el creciente uso de estas nuevas técnicas (APA, 1986). Esta expansión se explica por las ventajas de esta tecnología de evaluación con respecto a la tradicional de papel y lapiz.

En este capítulo se exponen inicialmente las aportaciones de los tests informatizados y la problemática derivada del uso del ordenador en la evaluación mediante tests. A continuación se presentan las características más importantes de los dos tipos más importantes de tests informatizados que se emplean en la actualidad: los tests informatizados basados en la Teoría Clasica de los Tests (TCTI) y los

tests adaptativos informatizados, basados en la Teoría de Respuesta al Item (TAI). Finalmente, se describen brevemente las lineas de desarrollo futuro.

2 - Aportaciones de los tests informatizados

Las aportaciones específicas del soporte informático en el ámbito de la evaluación psicométrica se puede concretar en los siguientes aspectos:

Economía:

Hay gente que piensa que la implementación de un sistema de tests informatizados es cara a causa del precio de los equipos. Sin embargo, los informes de las instituciones que emplean con frecuencia evaluaciones mediante tests (universidades, empresas, ejército, etc) permiten afirmar que el sistema informatizado reduce los costos a medio plazo (Backhoff, Ibarra y Rosas, 1997). Los tests de papel y lapiz pueden requerir mucho tiempo y personal en la aplicación, corrección, interpretación y almacenamiento de los datos. Los costos son mucho menores cuando se usan tests informatizados, puesto que todo el proceso está automatizado y sólo es necesario contar con algun supervisor para velar por el uso correcto del soporte. El ordenador evita que el profesional dedique mucho tiempo a tediosas tareas burocráticas.

Estandarización:

Una de las principales amenazas a la comparabilidad de las medidas es la variación introducida por los usuarios de los tests convencionales en las condiciones de aplicación (instrucciones, tiempo, etc) y corrección. En los tests informatizados las condiciones de aplicación y corrección son dirigidas rígidamente por un procedimiento automatizado, existiendo una absoluta garantía de que son iguales para todos los examinados.

Interacción con el examinado:

En los tests convencionales de aplicación colectiva, el examinado es dejado a su suerte. Una ventaja de los tests informatizados es que permiten dar un feedback inmediato al sujeto examinado para informarle de su grado de comprensión de las instrucciones (¡Comprueba que la respuesta correcta es la opción C!), de la corrección de su respuesta o de su balance rapidez-precisión (¡Trabaja más deprisa!). El feedback mejora la comprensión de la tarea y la motivación del sujeto y, en consecuencia, incrementa la validez (Kyllonen, 1991).

Seguridad:

Los sistemas informáticos permiten mejorar la seguridad del test ante robos, copias o usos no autorizados. Los ficheros que contienen la prueba pueden ser protegidos por claves de usuario que impidan el acceso, la copia o la impresión. Los ítems pueden ser presentados aleatoriamente para impedir la copia de examinados que ejecutan la prueba en puestos cercanos (Bunderson, Inouye, & Olsen, 1989). Además, se pueden administrar aleatoriamente formas paralelas de los tests TCTI para dificultar la transmisión del contenido del test a los sujetos examinados en turnos sucesivos.

Fiabilidad:

Para ahorrar material o permitir la corrección electrónica, en muchos tests convencionales se usan hojas en la que los sujetos marcan su respuesta a los ítems. Este procedimiento de registro de las respuestas es lento y ocasiona errores de anotación que decrementan la fiabilidad. Bounderson, Inouye y Olsen (1989) citan varios estudios en los que se demuestra empíricamente que las versiones informatizadas de los tests requieren menos tiempo de aplicación que las versiones en formato de papel y lapiz. Este ahorro de tiempo puede ser invertido en incluir un mayor número de ítems. Esta operación produce que la fiabilidad de los tests informatizados sea mayor que la de los tests de papel y lapiz. Como es sabido, la longitud del test está asociada a su fiabilidad, de forma que aumentando la longitud se obtendrá un aumento de la fiabilidad bajo ciertas condiciones (ausencia de violación del supuesto de paralelismo en la TCT e inclusión de los ítems más informativos en la TRI).

Riqueza del material estimular:

El ordenador permite construir y presentar ítems que no son asequibles al formato impreso. El software actualmente disponible permite una gran flexibilidad en la elaboración del material estimular utilizando textos, gráficos, animación, imágenes en movimiento, sonidos, video, audio, etc. Este abanico de posibilidades permite construir instrucciones de gran calidad para incrementar la comprensión de la tarea y la motivación del sujeto, y nuevos tipos de ítems que hacen posible nuevos modelos de tests y la medición de rasgos dificilmente evaluables mediante tests impresos (flexibilidad atencional, coordinación de distintas fuentes de información, memoria auditiva, percepción del movimiento, percepción del tiempo, etc).

Capacidad y rapidez de almacenamiento:

Los tests informatizados permiten el almacenamiento de los datos sin etapas previas de codificación y escritura. Además, hacen posible un rápido almacenamiento de una ingente cantidad de variables que describen la conducta del examinado durante el test (comprensión de las instrucciones y los ejemplos, opciones de respuesta seleccionadas, latencias de respuesta a cada ítem, ítems omitidos y no alcanzados, etc). El registro de variables temporales es especialmente importante porque permite medir aptitudes como la rapidez de procesamiento (Kyllonen, 1991) o rasgos de personalidad como la propensión a mentir (Neubauer & Malle, 1997) que son inasequibles para los tests de papel y lapiz. El registro de los tiempos de reacción es de suma importancia para el desarrollo de tests aptitudinales a partir del enfoque del procesamiento de la información (Ronning, Conoley, Glover, & Witt, 1987; Snow & Lohman, 1989). Desde esta nueva perspectiva de construcción de tests, se diseñan tests integrados por subconjuntos de items que suscitan el empleo de distintas estrategias de resolución de tareas (Carroll, 1987; Lohman & Ippel, 1993). Las latencias son las variables dependientes más usadas para contrastar los modelos de procesamiento. Las pruebas informatizadas permiten obtener y almacenar de forma rápida y precisa este tipo de datos.

Popham, W. J. (1971). Indices of adequacy for criterion-referenced tests items. In W. J. Popham (Ed.), *Criterion-referenced measurement.* Englewood Cliffs, NJ: Educational Technology Publications.

Popham, W. J. (1975). *Educational evaluation.* Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Shaycoft, M. F. (1979). *Handbook of criterion-referenced testing: Development, evaluation, and use.* New York: Garland STPM.

Swaminathan, H., Hambleton R. K. & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement, 11*, 262-267.

TESTS INFORMATIZADOS

Gerardo Prieto Adánez

1 - Introducción

El término *test informatizado* se emplea en la actualidad para describir un tipo de test en el que el ordenador es el soporte de todas las fases de la ejecución de la prueba:

(a) presentación en pantalla de las instrucciones
(b) presentación en pantalla de los ejemplos de práctica
(c) presentación en pantalla de los ítems
(d) registro de datos censales emitidos mediante el teclado o el ratón (nombre, edad, sexo, etc)
(e) registro de las respuestas emitidas mediante el teclado, el ratón, micrófono, o pantalla sensible al tacto
(f) codificación numérica de las respuestas
(g) almacenamiento de los datos
(h) puntuación de la prueba
(i) emisión de informes sobre los resultados.

Por tanto, no ha de aplicarse el calificativo de informatizado a los tests de papel y lapiz que se contestan en hojas de respuesta legibles mediante una lectora óptica conectada a un ordenador.

El origen y la proliferación de este tipo de tests es paralela a la difusión de los ordenadores personales. Desde la década de los 70 con las primeras versiones informatizadas del MMPI (Lushene, O'Neil, & Dunn, 1974), el crecimiento de esta metodología ha sido exponencial en la psicología y la educación. La publicación por la APA de una normativa para la administración e interpretación de los tests informatizados se ha debido a la importancia y el creciente uso de estas nuevas técnicas (APA, 1986). Esta expansión se explica por las ventajas de esta tecnología de evaluación con respecto a la tradicional de papel y lapiz.

En este capítulo se exponen inicialmente las aportaciones de los tests informatizados y la problemática derivada del uso del ordenador en la evaluación mediante tests. A continuación se presentan las características más importantes de los dos tipos más importantes de tests informatizados que se emplean en la actualidad: los tests informatizados basados en la Teoría Clasica de los Tests (TCTI) y los

tests adaptativos informatizados, basados en la Teoría de Respuesta al Item (TAI). Finalmente, se describen brevemente las lineas de desarrollo futuro.

2 - Aportaciones de los tests informatizados

Las aportaciones específicas del soporte informático en el ámbito de la evaluación psicométrica se puede concretar en los siguientes aspectos:

Economía:

Hay gente que piensa que la implementación de un sistema de tests informatizados es cara a causa del precio de los equipos. Sin embargo, los informes de las instituciones que emplean con frecuencia evaluaciones mediante tests (universidades, empresas, ejército, etc) permiten afirmar que el sistema informatizado reduce los costos a medio plazo (Backhoff, Ibarra y Rosas, 1997). Los tests de papel y lapiz pueden requerir mucho tiempo y personal en la aplicación, corrección, interpretación y almacenamiento de los datos. Los costos son mucho menores cuando se usan tests informatizados, puesto que todo el proceso está automatizado y sólo es necesario contar con algun supervisor para velar por el uso correcto del soporte. El ordenador evita que el profesional dedique mucho tiempo a tediosas tareas burocráticas.

Estandarización:

Una de las principales amenazas a la comparabilidad de las medidas es la variación introducida por los usuarios de los tests convencionales en las condiciones de aplicación (instrucciones, tiempo, etc) y corrección. En los tests informatizados las condiciones de aplicación y corrección son dirigidas rígidamente por un procedimiento automatizado, existiendo una absoluta garantía de que son iguales para todos los examinados.

Interacción con el examinado:

En los tests convencionales de aplicación colectiva, el examinado es dejado a su suerte. Una ventaja de los tests informatizados es que permiten dar un feedback inmediato al sujeto examinado para informarle de su grado de comprensión de las instrucciones (¡Comprueba que la respuesta correcta es la opción C!), de la corrección de su respuesta o de su balance rapidez-precisión (¡Trabaja más deprisa!). El feedback mejora la comprensión de la tarea y la motivación del sujeto y, en consecuencia, incrementa la validez (Kyllonen, 1991).

Seguridad:

Los sistemas informáticos permiten mejorar la seguridad del test ante robos, copias o usos no autorizados. Los ficheros que contienen la prueba pueden ser protegidos por claves de usuario que impidan el acceso, la copia o la impresión. Los ítems pueden ser presentados aleatoriamente para impedir la copia de examinados que ejecutan la prueba en puestos cercanos (Bunderson, Inouye, & Olsen, 1989). Además, se pueden administrar aleatoriamente formas paralelas de los tests TCTI para dificultar la transmisión del contenido del test a los sujetos examinados en turnos sucesivos.

Fiabilidad:

Para ahorrar material o permitir la corrección electrónica, en muchos tests convencionales se usan hojas en la que los sujetos marcan su respuesta a los ítems. Este procedimiento de registro de las respuestas es lento y ocasiona errores de anotación que decrementan la fiabilidad. Bounderson, Inouye y Olsen (1989) citan varios estudios en los que se demuestra empíricamente que las versiones informatizadas de los tests requieren menos tiempo de aplicación que las versiones en formato de papel y lapiz. Este ahorro de tiempo puede ser invertido en incluir un mayor número de ítems. Esta operación produce que la fiabilidad de los tests informatizados sea mayor que la de los tests de papel y lapiz. Como es sabido, la longitud del test está asociada a su fiabilidad, de forma que aumentando la longitud se obtendrá un aumento de la fiabilidad bajo ciertas condiciones (ausencia de violación del supuesto de paralelismo en la TCT e inclusión de los ítems más informativos en la TRI).

Riqueza del material estimular:

El ordenador permite construir y presentar ítems que no son asequibles al formato impreso. El software actualmente disponible permite una gran flexibilidad en la elaboración del material estimular utilizando textos, gráficos, animación, imágenes en movimiento, sonidos, video, audio, etc. Este abanico de posibilidades permite construir instrucciones de gran calidad para incrementar la comprensión de la tarea y la motivación del sujeto, y nuevos tipos de ítems que hacen posible nuevos modelos de tests y la medición de rasgos dificilmente evaluables mediante tests impresos (flexibilidad atencional, coordinación de distintas fuentes de información, memoria auditiva, percepción del movimiento, percepción del tiempo, etc).

Capacidad y rapidez de almacenamiento:

Los tests informatizados permiten el almacenamiento de los datos sin etapas previas de codificación y escritura. Además, hacen posible un rápido almacenamiento de una ingente cantidad de variables que describen la conducta del examinado durante el test (comprensión de las instrucciones y los ejemplos, opciones de respuesta seleccionadas, latencias de respuesta a cada ítem, ítems omitidos y no alcanzados, etc). El registro de variables temporales es especialmente importante porque permite medir aptitudes como la rapidez de procesamiento (Kyllonen, 1991) o rasgos de personalidad como la propensión a mentir (Neubauer & Malle, 1997) que son inasequibles para los tests de papel y lapiz. El registro de los tiempos de reacción es de suma importancia para el desarrollo de tests aptitudinales a partir del enfoque del procesamiento de la información (Ronning, Conoley, Glover, & Witt, 1987; Snow & Lohman, 1989). Desde esta nueva perspectiva de construcción de tests, se diseñan tests integrados por subconjuntos de items que suscitan el empleo de distintas estrategias de resolución de tareas (Carroll, 1987; Lohman & Ippel, 1993). Las latencias son las variables dependientes más usadas para contrastar los modelos de procesamiento. Las pruebas informatizadas permiten obtener y almacenar de forma rápida y precisa este tipo de datos.

Facilidad y rapidez de puntuación:

El sistema informatizado puede programarse para que, de forma inmediata, automática, estandarizada, rápida y precisa, puntúe los ítems ponderando las respuestas de la forma apropiada y para que combine las puntuaciones en los ítems a fin de obtener puntuaciones globales en diversas subescalas y variables (precisión, rapidez, etc). El software actualmente disponible permite la estimación de los parámetros de los ítems y de las puntuaciones en el rasgo en los tests basados en la TRI.

Obtención inmediata de informes estandarizados:

Además de efectuar los cálculos necesarios para puntuar el test una vez finalizado, los tests informatizados suelen ser programados para proporcionar informes del significado de las puntuaciones pocos segundos despues de terminar la prueba. Esta rapidez es una gran ventaja, puesto que muchos tests de papel y lapiz requieren un tiempo mucho mayor. La información emitida puede tener distintos grados de complejidad presentando sólo una simple descripción de las puntuaciones obtenidas o además un análisis de los resultados y un informe diagnóstico.

Los informes pueden ir dirigidos al evaluado o al evaluador. El primer caso es frecuente en los examenes académicos tipo test en los que se informa al examinado de su patrón de resultados (aciertos, errores, omisiones), de la calificación obtenida y de sus consecuencias (superación o no del examen). Estos informes pueden incluir eventualmente prescripciones educativas (identificación de conceptos mal comprendidos, recomendación de lecturas o ejercicios de práctica, etc).

Los informes dirigidos al evaluador suelen tener un carácter más técnico. Incluyen una descripción estadística de los resultados (transformación de las puntuaciones directas en baremos normativos, representación gráfica de perfiles, etc) y un informe diagnóstico estandarizado (Lopez, Ato, Sanchez & Velandrino, 1990). Un ejemplo muy conocido es la versión informatizada del MMPI desarrollada por Fowler (1987) en la que se proporciona un extenso informe dirigido al clínico en el que se incluyen los siguientes apartados: validez del perfil, patrón de síntomas, relaciones interpersonales, estabilidad de la conducta, consideraciones para el diagnóstico y consideraciones para el tratamiento.

3 - Interacción entre el evaluado y el ordenador

El empleo del ordenador para administrar tests ha generado una nueva problemática relacionada con la equivalencia de los diversos medios de evaluación y con los aspectos característicos de la evaluación informatizada. Los primeros tests informatizados fueron meras versiones «traducidas» de los tests de papel y lapiz. Es decir, las instrucciones y los ítems eran de un formato similar. En consecuencia, tenía sentido comparar los resultados psicométricos obtenidos con ambos tipos de soporte y tambien los efectos psicológicos en los evaluados. Los estudios sobre la equivalencia psicométrica indican que la fiabilidad es similar (Lopez et al., 1990). Sin embargo, existen diferencias entre las medias de las puntuaciones aportadas por tests con distinto soporte. Mazzeo y Harvey (1988), tras revisar un gran número de estudios de comparación, concluyen que las puntuaciones no son equivalentes en la

mayor parte de los casos: en general los sujetos obtienen mejores puntaciones en los tests impresos (explicables por la posibilidad de omitir o corregir las respuestas), aunque las diferencias suelen ser pequeñas y de escasa significación práctica. No obstante, se ha de tener en cuenta que la revisión de Mazzeo y Harvey incluye muchos tests informatizados desarrollados con escasos medios técnicos en las primeras etapas de desarrollo de esta metodología. Algunas de sus conclusiones han sido replicadas en posteriores investigaciones, como la de que el tamaño de las diferencias depende del contenido de los tests. Rolls y Feltham (1993) concluyen que las diferencias en tests de personalidad y actitudes son escasas o nulas. Por el contrario Brand y Houx (1992) y Bartram (1993) consideran que la equivalencia es menor en los tests de aptitudes, especialmente en los tests en los que es importante la velocidad de ejecución, dado que el teclado o el ratón permiten una mayor rapidez de respuesta.

En otros trabajos se ha analizado la influencia del soporte informático en la ansiedad de los sujetos, la sinceridad de sus respuestas, la motivación, las actitudes, etc. Las personas que carecen de experiencia con los ordenadores pueden tener prejuicios hacia los tests informatizados empleados en contextos de selección o evaluación académica. La suposición de que no tienen la competencia requerida para una ejecución correcta y el miedo a un procedimiento desconocido podría producirles actitudes negativas e incrementarles la ansiedad en la situación de prueba. No obstante, se han obtenido resultados contradictorios en las investigaciones que han contrastado empíricamente estas hipótesis. Por ejemplo, Wise, Barnes, Harvey y Plake (1989) no encontraron diferencias en ansiedad ante los ordenadores entre dos grupos que respondieron a versiones informatizadas e impresas de una prueba de matemáticas. Schmitt, Gilliland, Landis y Devine (1993) concluyeron que los aspirantes a puestos administrativos manifestaban actitudes más positivas y niveles mayores de motivación ante los tests informatizados. Schoonman (1989) encontró asimismo actitudes más positivas hacia los tests informatizados entre los aspirantes a puestos en los ferrocarriles holandeses. Por el contrario Held, O'Neil y Hansen (1973) encontraron un nivel más alto de ansiedad en las aplicaciones informatizadas del WAIS. Consideramos que se puede modificar la negatividad con que se percibe la situación mediante el uso de software de entrenamiento, instrucciones apropiadas para los tests, un número elevado de ejemplos con feedback y el diseño del test para que se puedan revisar las instrucciones, los ejemplos y las respuestas, y corregir las contestaciones previas.

En el contesto de la evaluación clínica y psiquiatrica, existen evidencias de que el formato informatizado facilita la sinceridad de las respuestas a preguntas íntimas (Evan & Miller, 1969), a preguntas relacionadas con conductas socialmente desviadas (Plutchik & Karasu, 1991) y con el consumo de alcohol (Schmitt, Gilliland, Landis, & Devine, 1993).

En cualquier caso, parece claro que las diferencias introducidas por el soporte aconsejan llevar a cabo análisis de la equivalencia de las puntuaciones de versiones impresas e informatizadas del mismo test. Las puntuaciones pueden ser consideradas equivalentes cuando (a) el orden de las puntuaciones de los sujetos es aproximadamente el mismo en ambas versiones, y (b) las medias, varianzas y formas de las distribuciones de puntuaciones son aproximadamente iguales. Si no se cumple la condición (a), las versiones no son equivalentes y, por tanto, no pueden ser usadas

de forma intercambiables y deben ser baremadas por separado. Si se cumple la condición (a) pero no la (b), es necesario transformar las puntuaciones a la misma escala para obtener dos versiones equivalentes e intercambiables (APA, 1986).

A partir de las dos últimas décadas está creciendo el número de las investigaciones sobre la interacción entre el usuario y el ordenador buscando determinar las condiciones óptimas que ha de tener el sistema informático para adecuarse a las características específicas de los usuarios (nivel cultural, competencia en el uso de ordenadores, presencia de discapacidades sensoriales y motoras, etc). Esta temática suele cobijarse bajo el rótulo de *Factores Humanos*. Son muy diversos los objetivos de los trabajos (impacto de los ordenadores en las organizaciones, diseño de hardware y software interactivo, diseño de tareas, análisis de los modos de enfrentarse al ordenador, etc) y los campos científicos implicados (lingüística computacional, inteligencia artificial, ciencia cognitiva, sociologia, ergonomía, psicologia social, matemáticas, psicología cognitiva, ingeniería de sistemas, etc). Aunque esta variedad temática tiene un impacto diverso en la construcción de los tests informatizados, no puede ser desarrollada en el corto espacio de este capitulo. El lector interesado puede consultar el excelente libro de Booth (1989).

No obstante, mencionaremos algunas prescripciones relativas al hardware y al software que son importantes para la construcción y el uso de los tests informatizados.

Debido a los avances actuales en la tecnología informática, la mayor parte de los ordenadores personales disponibles en el mercado pueden ser usados como soporte de los tests informatizados. El equipo básico podría consistir en un ordenador personal con un procesador de 16 bit, co-procesador aritmético, disco duro y pantalla con alta resolución. Si los tests van a ser administrados periodicamente a ingentes cantidades de sujetos, es más aconsejable disponer de un sistema más complejo con un ordenador principal, que controla el desarrollo de la aplicación, la comunicación, el almacenamiento y el análisis de los datos, y varios ordenadores, autónomos y conectados al ordenador principal, que sirven como puestos de aplicación de tests (Hambleton, Zaal, & Pieters, 1991). Un problema central de los tests informatizados es la presentación de gráficos de alta calidad, los cuales requieren tanto una buena resolución de la pantalla como una elevada capacidad de memoria. El almacenamiento de los gráficos mediante CD-ROM y el uso de programas multimedia como el SuperCard (Allegiant, 1996) y MetaCard ha eliminado este problema en los últimos años.

El software más apropiado es el que permite incluir todos los componentes del proceso: construcción de los ítems y las instrucciones, presentación del test, almacenamiento de los datos, análisis de los datos, calibración de los ítems, puntuación del test y emisión de los informes (Hambleton, Zaal, & Pieters, 1991). El software idóneo debería ser lo suficientemente flexible para incorporar nuevos desarrollos sin tener que comenzar de cero. Probablemente el software comercial que mejor se adapta a estas especificaciones es el paquete MicroCAT 3.5 (Assessment Systems Corporation, 1994), aunque no existen versiones del programa para el entorno Windows ni para el entorno Macintosh. No obstante, dado el rápido desarrollo experimentado por el software, es muy posible que hayan aparecido nuevas ofertas cuando estas páginas salgan a la luz.

Bajo la denominación de *Factores Humanos*, se han formulado otras recomendaciones para los constructores de tests informatizados dirigidas al diseño de la información presentada por el sistema informático y a la aplicación de los tests. A continuación, se presenta una lista de las más importantes:

(a) Los tests informatizados deben permitir a los evaluados registrar sus respuestas de forma sencilla y confortable (mediante ratón, pantalla sensible al tacto, etc), comprobarlas y corregirlas si lo consideran necesario (APA, 1985)
(b) Los evaluados deberían ser claramente informados de los factores que son importantes para su ejecución: énfasis en la rapidez o la precisión, feedback sobre las respuestas a los ejemplos de práctica, procedimiento de puntuación, etc (APA, 1985)
(c) El entorno del terminal deberá ser tranquilo, confortable e impedir las distracciones (Hambleton, Zaal, & Pieters, 1991)
(d) La pantalla debería ser colocada de forma que no sea afectada por reflejos (Hambleton, Zaal, & Pieters, 1991)
(e) La legilibilidad de la pantalla debería ser evaluada empíricamente (Hambleton, Zaal y Pieters, 1991)
(f) Los ordenadores deben estar separados por pantallas que dificulten la visión de los que trabajan en puestos contigüos (Green, 1990)
(g) El administrador debe de tener espacio suficiente para moverse libremente entre los examinados y para facilitarles ayuda en caso necesario (Green, 1990)
(h) El intervalo de tiempo entre la presentación de ítems sucesívos debe de ser pequeño: entre 5 y 10 segundos (Green, 1990)
(i) Entre la desaparición de un ítem y la aparición del siguiente, la pantalla debe de mantener un color que no deslumbre (Green, 1990)
(j) El contenido del ítem (especialmente las gráficas) aparecerá y desaparecerá instantáneamente sin que se observe una aparición gradual en su presentación (Green, 1990)
(k) Se deben elaborar instrucciones específicas para la presentación informatizada (no meras adaptaciones de las instrucciones de los tests de papel y lapiz). El error más comun es llenar de texto la pantalla. Es preferible presentar una sóla idea o un sólo concepto en una ventana. Es interesante usar ventanas diferentes y simultáneas para las instrucciones y los ítems (Kyllonen, 1991)
(l) Se dede comenzar con un ítem, no con la teoría (informar exactamente al sujeto sobre lo que tiene que hacer). Cualquier mensaje teórico (la aptitud que mide el test) puede interferir la ejecución (Kyllonen, 1991)
(m) Es conveniente usar animación en las instrucciones de tests con ítems gráficos (Prieto & Delgado, 1996)
(n) Se debe proporcionar feedback sobre la rapidez y la precisión al término de cada ítem o de subconjuntos homogéneos de ítems (Kyllonen, 1991)
(o) Conviene informar de la teoría subyacente al terminar el test: rasgo medido, importancia en la vida diaria, etc (Kyllonen, 1991).

4 - Tests informatizados convencionales

Los tests informatizados convencionales constituyen la gran mayoría de las pruebas informatizadas que se aplican en la actualidad. Bunderson, Inouye y Olsen (1989) los clasifican dentro de la *primera generación* de esta tecnología en la que buena parte de los tests informatizados no difieren en sus aspectos más básicos de los tests de papel y lapiz (modelo matemático, contenido, puntuación y baremación, etc). Las diferencias radican en la optimización de los procesos aportadas por el soporte informático que ya se han comentado en un apartado anterior.

Puesto que la metodología fundamental es similar a la de los tests impresos, hemos calificado a estas pruebas de *convencionales* y dedicaremos poco espacio a describir sus características. Las principales son las que siguen:

(a) Se fundamentan en la Teoría Clásica de los Tests (TCT) que es el modelo matemático usado para medir la fiabilidad en la casi totalidad de los tests actuales. El modelo TCT permite evaluar la fiabilidad de los tests mediante el coeficiente de fiabilidad y el error típico de medida. El coeficiente de fiabilidad se define como la correlación entre tests paralelos y el error típico de medida se calcula a partir del coeficiente de fiabilidad y de la desviación típica de las puntuaciones observadas $\left(\sigma_e = \sigma_X\sqrt{1-\rho_{XX}}\right)$. Los procedimientos empíricos para calcular estos estadísticos son diseños de recogida de datos que intentan obtener variables que se ajusten a la definición de tests paralelos (formas paralelas, test-retest, dos mitades y coeficiente a de Cronbach). La homocedasticidad del error típico de medida es un supuesto clave y poco plausible de la TCT. Significa que el error tiene el mismo tamaño a lo largo de la variable medida. Con frecuencia los datos no se ajustan al supuesto (se mide con menos precisión en los extremos del continuo).

(b) Los examinados reciben los mismos ítems y en el mismo orden. Esta estandarización es una condición imprescindible para que las puntuaciones de distintos sujetos estén en la misma escala (sean comparables).

(c) Los procedimientos de puntuación se basan en la acumulación de puntos (suma de los valores obtenidos en los ítems). Se asume que los ítems son invariantes (valen lo mismo): acertar un ítem suma un punto, sea el ítem facil o dificil. Terminada la presentación de los ítems, el sistema informático puntúa el test, almacena y presenta el resultado.

(d) Uso de normas de grupo. Las puntuaciones de los tests TCT no son directamente relacionables con el nivel en el rasgo ni son equiparables directamente con las de otros tests. Por ello, es necesario elaborar normas cuantitativas que permitan interpretar el nivel de los sujetos. La mayor parte de los tests TCT usan *normas de grupo,* denominadas así porque se basan en una comparación de la puntuación de un sujeto con la distribución de puntuaciones en el grupo al que pertenece (población). La distribución de las puntuaciones en la población se estima mediante una muestra representativa denominada *grupo normativo* . La mayor parte de las normas se presentan mediante transformación de las puntuaciones directas en escalas como los percentiles (porcentaje de casos del grupo normativo que puntuan por debajo de dicha puntuación), puntuaciones típicas normalizadas (z) y puntuaciones típicas normalizadas derivadas (eneatipos, Cociente Intelectual de desviación, etc). Los tests informatizados son programados para presentar esta información de forma inmediata.

(e) Incorporan algunas novedades como el registro de variables temporales. En muchos casos, se programa el sistema para que contabilice las latencias a los ítems. Estas variables, que no puede ser registradas en los tests de papel y lapiz, se emplean de forma generalizada en la construcción de tests desde el paradigma de la psicología cognitiva (Lohman & Ippel, 1993) y como indicador de algunos constructos de personalidad (Neubauer & Malle, 1997).

Actualmente existe más de un millar de tests informatizados convencionales que miden los más diversos atributos en la psicología y la educación (Sweetland & Keyser, 1991).

5 - Tests adaptativos informatizados

5.1 - Introducción

Como hemos indicado en el apartado anterior, tanto en los tests de papel y lapiz como en los informatizados convencionales se presentan al examinado todos los ítems que componen el test y en un orden estandarizado. Este procedimiento es poco económico y eficiente. Supongamos que un test de Comprensión Verbal, construido de acuerdo con el modelo TCT, se administra a una submuestra de sujetos de muy alto nivel. Obviamente, todos los sujetos resolverán correctamente los ítems de dificultad media y baja, y sólo diferirán entre sí en las respuestas a los ítems de dificultad elevada. En consecuencia, la aplicación de la mayor parte del test es ineficaz e innecesaria para evaluar el nivel de los más aptos. Se puede formular el mismo juicio negativo sobre la adecuación de los ítems medianamente difíciles y muy difíciles para evaluar a los menos capacitados. Una solución al problema consistiría en adaptar la dificultad del ítem al nivel del examinado en el rasgo. Con esta estrategia se conseguiría una gran rapidez en la evaluación (hay que administrar pocos ítems) y una gran precisión (sólo se administran los ítems imprescindibles para concretar la situación del sujeto en la dimensión medida). Además, se incrementaría la seguridad del test, puesto que los sujetos sólo pueden conocer una submuestra de ítems.

Hay que resaltar que este procedimiento de evaluación requiere un modelo de test distinto de la TCT porque:

(a) El procedimiento de puntuación no puede basarse en el número de ítems resueltos correctamente (sería igual la puntuación de un sujeto que resuelve cinco ítems fáciles que la del que resuelve cinco ítems difíciles)

(b) Las puntuaciones obtenidas tras contestar a distintos bloques de ítems deben estar en la misma escala

(c) Se debe emplear un procedimiento dinámico de evaluación que permita estimar el nivel aproximado del sujeto antes de seleccionar el ítem o los ítems que se le van a administrar.

Este método de evaluación adaptada al sujeto se viene empleando de manera informatizada en los últimos 20 años, de ahí el término de *test adaptativo informatizado* (computerized adaptive testing, en su denominación inglesa). Los tests adaptativos informatizados (TAI) requieren el uso de la Teoría de Respuesta al Item

(TRI), siendo en realidad la aplicación estrella de estos nuevos modelos de tests. Por ello, resumiremos algunas nociones básicas de TRI, imprescindibles para seguir con provecho la exposición de los TAI.

5.2 - Teoría de la Respuesta al Item (TRI).

Desde comienzos de siglo, la construcción y el uso de los tests se ha fundamentado mayoritariamente en la TCT y, debido a las escasas restricciones que impone a los datos, su vigencia puede durar mucho tiempo. Sin embargo, existen varias limitaciones teóricas y prácticas de la TCT: (a) la crucial definición de tests paralelos no es verificable empíricamente, (b) se suele asumir inadecuadamente que el error típico de medida es el mismo para todos los niveles de la variable (no es cierto que se mida con el mismo error en el centro que en los extremos), (c) las estimaciones del nivel de un sujeto procedentes de distintos tests son distintas y dificilmente equiparables[1], y (d) las propiedades psicométricas del test dependen de las muestras de sujetos analizadas[2].

Estas limitaciones explican la aparición de nuevos modelos de construcción de tests que se enmarcan en la TRI. Aunque los orígenes de la TRI se sitúan en la década de los treinta (Muñiz & Hambleton, 1992), su expansión se ha producido a partir de los años ochenta debido a la difusión de los ordenadores, cuya utilización en este campo es imprescindible. De la extensa bibliografía sobre el tema, se destacan algunos títulos muy relevantes como los publicados por Lord (1980), Hambleton y Swaminathan (1985), Hambleton, Swaminathan y Rogers (1991) y Van der Linden y Hambleton (1996).

Los objetivos básicos de la TRI son dos:

(a) Obtener estimaciones del nivel de los sujetos en el rasgo que no dependan de la muestra de ítems que se les administre. Es decir, la puntuación de un sujeto habrá de ser la misma, sea cual sea el subconjunto de ítems al que responda

(b) Construir tests cuyas características psicométricas no dependan de la muestra de sujetos. Ello significa que las propiedades del test no varian de muestra a muestra.

Cuando los datos empíricos se ajustan a los restrictivos supuestos de algun modelo TRI, ambos objetivos se cumplen.

El concepto básico de la TRI es la Curva Característica del Item (CCI) que es la función matemática que relaciona la probabilidad de acertar el ítem[3] con la competencia de los sujetos (?). Se denomina CCI porque cada ítem tiene su particular curva. Esta es una clara diferencia de la TRI respecto a la TCT. Mientras que la TCT va dirigida a las propiedades de la puntuación global en un test (suma de los valores de un conjunto de ítems), la TRI se centra en las propiedades particulares de cada ítem. Está, por tanto mas cercana a la explicación de la conducta (que se manifiesta en la respuesta a ítems concretos). Cuando se suman o promedian ítems para obtener una puntuación global (procedimiento clásico de puntuación de los tests TCT), la representación de la conducta del sujeto está más difuminada.

En la Figura 9-1 aparecen ejemplos de distintas CCI.

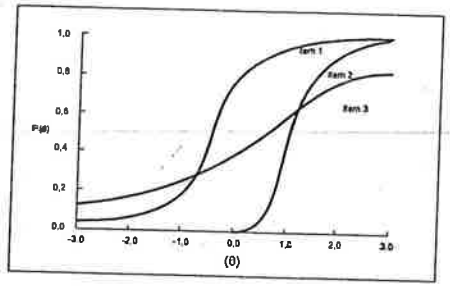


Fig.9-1. CCI de tres ítems

Como puede verse en la Figura 9-1, las CCI se representan en unos ejes de coordenadas, situándose en la ordenada la probabilidad de acertar el ítem (P(θ)) y en la abscisa los valores de la variable medida por los ítems (θ), cuyo rango puede oscilar desde + ∞ a - ∞. Las curvas suelen reflejar una relación monotónica creciente indicativa de que la probabilidad de acertar es mayor a medida que aumenta el nivel en el rasgo. La CCI puede ser caracterizada por una serie de parámetros. Los más usados, identificados por las tres primeras letras del alfabeto, son:

(1) el parámetro de discriminación (a) que es un valor proporcional a la pendiente de la recta tangente a la CCI en el punto de máxima pendiente. Cuanto mayor es *a*, menor será el cambio que se requiere en θ para que se incremente notablemente la probabilidad de acertar el ítem. En la Figura 9-1 se puede observar que 2 es el ítem más discriminativo y 3 el menos discriminativo. En el primer caso, es muy grande la diferencia en la probabilidad de acertar entre los sujetos con θ = 0 y θ = 1, mientras que en el segundo caso la diferencia es mucho menor. Es decir, el ítem 2 discrimina muy bien entre los sujetos con θ = 0 y θ = 1, el ítem 3 discrimina mucho peor entre los mismos sujetos. Nótese que un ítem tiene distinta eficacia discriminativa en distintos puntos del rasgo (el ítem 1 discrimina bien en el rango de θ = -2 a θ = 0 y no discrimina entre θ = 2 y θ = 0)

(2) el parámetro de dificultad (b) que es el valor de θ correspondiente al punto de máxima pendiente de la CCI. En los ejemplos de la Figura 9-1, se puede constatar que el ítem 2 (b = 1) es más dificil que el item 1 (b = -1). Obsérvese que los sujetos con θ = 1 tienen una alta probabilidad de acertar el item 2, mientras que los sujetos con θ = 0 tienen una probabilidad nula. Por el contrario, los sujetos con θ = 0 tienen una alta probabilidad de acertar el ítem 1. Una característica interesante de la TRI es que la dificultad de los ítems se mide en la misma escala del rasgo (en la misma que los sujetos). Por tanto, los modelos TRI se clasifican en el *enfoque centrado de las respuestas,* de acuerdo con la taxonomía de

1 La ausencia de invarianza de las mediciones respecto del intrumento utilizado plantea dificultades en los estudios evolutivos y serios problemas éticos en las evaluaciones educativas. ¿Como analizar la evolución de la inteligencia en un amplio rango del desarrollo si se emplea un test diferente para cada intervalo de edades? ¿Como comparar el nivel de los sujetos si se emplean distintos tests o exámenes?

2 Por ejemplo, la dificultad y la fiabilidad de un test pueden ser muy distintas en unas muestras y otras. ¿Cambian las propiedades del metro en función del objeto medido?

3 Acertar o fallar son palabras restringidas a los ítems de ejecución máxima. Ello no quiere decir que los modelos TRI no sean aplicables a los tests de ejecución típica (personalidad, actitudes, etc). Los vocablos acierto y competencia son lastres terminológicos introducidos por la aplicación original de la TRI: la evaluación académica.

Torgerson (1958). Se trata de una diferencia notable con la TCT. Los tests TCT se clasifican en el *enfoque centrado en los sujetos*, lo cuál supone que los ítems son invariantes (réplicas unos de otros); es decir, son indicadores similares del atributo medido (por eso el procedimiento de puntuación en los tests clásicos es la suma de los ítems). Por el contrario, los ítems de los tests TRI pueden tener distinto valor, reflejan distinto nivel en el constructo (la puntuación en el test no es el número de ítems acertados, depende del valor de los ítems que se acierten)

(3) el parámetro de *pseudoazar* (c) que es la probabilidad de acertar el ítem de los sujetos con muy bajo nivel en el rasgo.

El tipo de CCI que se adopte para el conjunto de ítems que habrán de componer los tests determinan el modelo TRI elegido. Los modelos TRI más usuales en la actualidad son los llamados modelos logísticos, debido a que se emplea la Función Logística para describir la CCI. Se usan fundamentalmente tres tipos: el modelo de un parámetro (los ítems sólo varían el parámetro *b*, siendo $c = 0$ y *a* constante), el modelo de dos parámetros (los ítems pueden variar en los parámetros *a* y *b*, siendo $c = 0$) y el modelo de tres parámetros (los ítems pueden variar en los parámetros *a*, *b* y *c*). A continuación, se describen las ecuaciones matemáticas correspondientes a los tres modelos (para una revisión, véase Hambleton, Swaminathan, & Rogers, 1991; Muñiz, 1996 y Pasquali, 1996).

Modelo logístico de un parámetro: $$P(\theta) = \frac{e^{D(\theta - b)}}{1 + e^{D(\theta - b)}}$$

Modelo logístico de dos parámetros: $$P(\theta) = \frac{e^{Da(\theta - b)}}{1 + e^{Da(\theta - b)}}$$

Modelo logístico de tres parámetros: $$P(\theta) = c + (1 - c)\frac{e^{Da(\theta - b)}}{1 + e^{Da(\theta - b)}}$$

siendo *a*, *b*, *c*, $P(\theta)$ y θ los valores antes definidos, *e* la base de los logaritmos neperianos (2,7182) y D una constante (cuando $D = 1,7$ la función logística es similar a la función normal).

Además de los supuestos específicos de cada modelo (relativos al tipo de CCI admisible), todos comparten algunos supuestos básicos: unidimensionalidad (los ítems y los sujetos tienen valores en una única dimensión o rasgo, de forma que la respuesta al ítem está determinada fundamentalmente por el nivel de los sujetos en la variable medida) e independencia local (en los sujetos con el mismo nivel en el rasgo, la respuesta a un ítem no está asociada con las respuestas a los demás ítems). Los modelos descritos se aplican a ítems dicotómicos en los que las respuestas pueden ser codificadas de forma binaria (correcto/incorrecto, verdadero/falso, acuerdo/desacuerdo, etc). Estas características de los modelos más usuales restringen su aplicación a un cierto tipo de tests. Como se comenta al final de este capítulo, en los últimos años se está investigando con modelos más complejos (multidimensionales, ítems politómicos, etc) que permitirán extender la TRI a otros campos como la medición de las actitudes, personalidad, etc (Van der Linden & Hambleton, 1996).

Además de los parámetros de la CCI previamente definidos, un ítem está caracterizado por su función de información ($I_i(\theta)$) que indica la fiabilidad con la que mide a lo largo del rasgo:

$I_i(\theta) = [P'(\theta)]^2 / P(\theta)\,Q(\theta)$, siendo $P(\theta)$ la probabilidad de acertar el ítem dado por la CCI, $Q(\theta) = 1 - P(\theta)$ y $P'(\theta)$ la derivada de $P(\theta)$.

Este indicador pone de manifiesto los niveles del atributo en los que el ítem es preciso y, en consecuencia, útil. Este aspecto es una de las características distintivas de la TRI. Conocer los rangos del atributo en los que un ítem es fiable o informativo constituye el criterio de selección de ítems mas usado en los tests adaptativos informatizados (Wainer, 1990).

5.3 - Análisis y Calibración del Banco de Items

La primera etapa de la construcción de un TAI consiste en *calibrar* un banco de ítems. Un banco de ítems es un conjunto de ítems que miden el mismo rasgo, tienen idéntico formato, se ajustan al mismo modelo TRI y tienen estimados en la misma escala sus valores psicométricos (parámetros de la CCI y función de información). El conjunto inicial de ítems que se va a utilizar para construir el banco puede provenir de tests ya existentes o ser construidos expresamente. Aunque el número de ítems del banco depende del constructo que se va a medir (hay atributos que no requieren muchos ítems), se suele aconsejar que el banco de un buen TAI debe contener al menos 100. Puesto que algunos ítems serán rechazados en el proceso de calibración, el conjunto inicial de ítems habrá de ser mayor.

Las fases del proceso de calibración son tres (Barbero, 1996): (a) administración de los ítems a una muestra de sujetos, (b) análisis del ajuste de los ítems a un modelo TRI y (c) estimación de los parámetros de los ítems y de su función de información.

a) Selección de la muestra y aplicación del conjunto inicial de ítems.

Para llevar a cabo una buena calibración, es conveniente garantizar que el test va a ser administrado a sujetos característicos de todos los niveles en el rasgo medido y de los distintos sectores de la población con diferentes niveles en dicho rasgo (sexo, edad, clase social, nivel educativo, etc). Aunque las técnicas de muestreo probabilísticas son las únicas que permiten garantizar la representatividad, no son utilizadas en muchos trabajos debido a su elevado costo y a la rutina. En la mayor parte de los casos, las muestras se eligen empleando algunos criterios racionales tendentes a garantizar que la muestra tiene características similares a la población. En este caso, se identifican algunas características relevantes de la población (sexo, edad, clase social, nivel educativo, etc) y, en base a ellas, se establece el tipo y número de sujetos a los que hay que administrar el test. Es deseable que el número de sujetos sea elevado, no inferior a 500 (Bunderson et al., 1989).

Cuando el número de ítems iniciales no es demasiado grande, es posible administrar todos ellos a la muestra de sujetos. Es conveniente administrar los ítems mediante un procedimiento informatizado convencional puesto que el soporte puede tener influencia en las características psicométricas de los ítems. Para evitar errores, se aconseja familiarizar al examinado con el procedimiento incluyendo ejemplos con feedback y varios ítems de prueba con características similares al resto del banco.

Si el número de ítems iniciales es muy grande, será practicamente imposible que todos los sujetos respondan a todos los ítems en una única sesión. La posible alternativa de administrar los ítems en varias sesiones es frecuentemente descartada a causa de los costos y de la pérdida de sujetos. La mejor solución de las posibles es emplear *un diseño de anclaje*. Este diseño consiste en dividir la muestra en subgrupos y aplicar a cada subgrupo un conjunto de ítems específicos y otro de ítems comunes, denominados *items de anclaje*. La mayor parte de los estudios manifiestan que no es necesario un número muy elevado de ítems de anclaje (Wingersky & Lord, 1984): entre 5 y 15 ítems serían suficientes.

b) Análisis del ajuste de los ítems a un modelo TRI.

El ajuste de los datos al modelo comienza por verificar si los ítems son unidimensionales. No suele efectuarse un contraste directo del supuesto de independencia local, puesto que de la unidimensionalidad se sigue necesariamente esta característica. Para verificar la unidimensionalidad se realiza un análisis factorial de las correlaciones tetracóricas entre los ítems. La matriz de correlaciones tetracóricas puede obtenerse previamente mediante el módulo PRELIS del SPSS (SPSS, Inc., 1994). Si el porcentaje de varianza explicada por el primer factor es elevado y muy superior al explicado por el segundo, se considera que los ítems son fundamentalemente unidimensionales. Para lograr este objetivo, se suele seguir un proceso de descarte de aquellos ítems que saturan en factores residuales hasta conseguir que un factor explique la mayor parte de la varianza (Muñiz, 1997).

Una vez seleccionados los ítems adecuados a la unidimensionalidad, se procede a verificar qué ítems se ajustan a los requerimientos particulares del modelo TRI seleccionado (de uno, dos o tres parámetros). Si, por ejemplo, se pretende construir un test con el modelo TRI más parsimonioso y restrictivo en el que se asume que el parámetro *a* es constante y el parámetro *c* igual a cero (el modelo de un parámetro), se habrá de analizar el ajuste de los ítems a esos supuestos. Inicialmente se estiman, partiendo de los supuestos anteriores, el parámetro de dificultad (b) de los ítems, el nivel de competencia de los sujetos (θ) y las probabilidades de acertar el ítem en distintos niveles de competencia ($P'(\theta)$). El procedimiento de contraste consiste en determinar si son significativas las diferencias entre las probabilidades de acertar el ítem observadas ($P(\theta)$) y estimadas desde el modelo ($P'(\theta)$). Como estadístico de contraste suele usarse c^2 en un procedimiento propuesto por Yen (1981). Si el estadístico no es significativo, se considera que el ítem se ajusta al modelo propuesto. Existen otros procedimientos de contraste del ajuste de los datos al modelo. Una excelente revisión puede encontrarse en Muñiz (1997).

Si se sospecha que los ítems pueden presentar funcionamiento diferencial (DIF) en distintos grupos sociales, es conveniente analizar este aspecto y descartar los ítems que presenten DIF. Un ítem tiene DIF (Differential Item Functioning) cuando la proporción de sujetos con el mismo nivel en el rasgo que resuelve correctamente el ítem presenta variaciones entre grupos (varones o mujeres, distintas culturas, etc). Se debe excluir este tipo de items para que los instrumentos de evaluación tengan la misma validez en los distintos grupos y, sobre todo, para evitar el sesgo del test. El sesgo se produce cuando el funcionamiento diferencial

ocasiona graves consecuencias para un subgrupo determinado de personas (Cole & Moss, 1989).

c) Estimación de los parámetros de los ítems y de su función de información.

Una vez seleccionados los ítems apropiados, se estiman sus características psicométricas (los parámetros de la CCI correspondientes al modelo elegido y la función de información). Existen varios programas de ordenador para analizar el ajuste de los datos al modelo, calcular las funciones de información y estimar los parámetros de los ítems (Hambleton, Swaminathan & Rogers, 1991). Los más utlizados son el BILOG (Mislevy & Bock, 1984) y los módulos RASCAL y ASCAL del MicroCAT (Assessment Systems Corporation, 1994). En el caso de haber utilizado un diseño de anclaje en la recogida de datos, se estimarán los parámetros de forma separada en cada una de las submuestras. Las ecuaciones de regresión que relacionan los parámetros de los ítems de anclaje en los distintos subgrupos permitirán situar los parámetros de todos los ítems en la misma escala (Barbero, 1996).

5.4 - Diseño de la estrategia adaptativa

Una vez calibrado el banco de ítems, se procede a diseñar la estrategia adaptativa que consiste en establecer los criterios de comienzo, presentación de ítems y terminación del test. De las diversas modalidades de estrategia adaptativa aquí presentaremos las tres más usuales: dos niveles, nivel flexible y máxima información.

a) Estrategia de dos niveles.

Se comienza la prueba presentando un subtest (denominado ***de rutina***) que consiste en una prueba breve con ítems de dificultad heterogénea. Mediante la rutina se estima un valor inicial (θ^*) del sujeto que determina la dificultad del subtest *localizador* que se le presentará a continuación. Para ello, se emplea alguno de los procedimientos derivados de la TRI: estimación de máxima verosimilitud o bayesiana (Hambleton & Swaminathan, 1985). El localizador es un subtest con ítems de dificultad homogénea. Se construyen varios localizadores de distinta dificultad promedio. Se administra al sujeto uno de ellos en función de la θ^* estimada en la rutina. A partir de las respuestas al localizador (y eventualmente a la rutina) se estima la puntuación θ' definitiva del sujeto. El éxito radica en la precisión con que la rutina estime θ. Esta estrategia, la menos adaptativa de todas, tiene su origen en las primeras versiones de los tests adaptativos aptas para su uso con tests de papel y lapiz (Lord, 1980). En aplicaciones informatizadas masivas no es muy adecuada por razones de seguridad (transferencia de información sobre el subtest de rutina a los turnos sucesivos de examinados).

b) Estrategia de nivel flexible.

Se inicia el test con la presentación de un ítem de dificultad media. Si el sujeto lo resuelve correctamente, se le presenta el ítem del banco que le sigue en

dificultad. Si la respuesta es incorrecta, se le presenta un ítem algo menos difícil. Se sigue una secuencia similar hasta que el sujeto ha contestado a un número de ítems predeterminado (suficiente para conseguir una precisión adecuada). Para evitar que siempre se presente el mismo ítem al comenzar, el sistema puede seleccionar aleatoriamente alguno entre un conjunto de ítems de dificultad media. Al término de la presentación, se estima el nivel de habilidad del sujeto (θ') mediante alguno de los procedimientos antes mencionados. Esta estrategia es especialmente apropiada para los casos en que se dispone de bancos con un reducido número de ítems (Olea & Ponsoda, 1996). Su deficiencias fundamentales radican en que: (a) la selección de un ítem se basa sólo en la respuesta al previo, no en la estimación del nivel del sujeto, y (b) puede existir cierta variabilidad entre los sujetos en la precisión de θ' (Hambleton, Zaal, & Pieters, 1991).

c) Estrategia de máxima información.

Es la que tiene mayor grado de adaptabilidad, dado que en el transcurso de la presentación se seleccionan los ítems más adecuados a la habilidad del sujeto estimada a partir de las respuestas emitidas hasta ese momento.

Se puede comenzar el test asignando a todos los evaluados un nivel medio de competencia ($\theta = 0$) o, para evitar que siempre se presente el mismo ítem, un valor seleccionado aleatoriamente de un rango medio de competencia (entre $\theta = +1$ y $\theta = -1$). Tras la respuesta del sujeto, se estima un nuevo nivel de habilidad (θ') y se presenta el ítem más fiable para ese nivel, es decir el que aporta mayor cantidad de información ($I_i(\theta)$). Los ítems más *informativos* para un nivel de habilidad son los que tienen un parámetro de dificultad similar ($b \approx \theta$), elevada discriminación (a) y bajo pseudoazar (c). Tras la nueva respuesta, se estima de nuevo la habilidad y se selecciona el ítem más informativo para el nuevo valor. El proceso continúa hasta que se consigue una precisión determinada previamente, por lo que no es necesario que todos los sujetos contesten a un mismo número de ítems.

La estrategia de máxima información es la más eficiente, puesto que permite conseguir un buen nivel de precisión con un pequeño número de ítems (Renom, 1993).

Existen diversas variantes de las estrategias adaptativas comentadas aquí. Una revisión exhaustiva puede encontrarse en el capítulo de Olea y Ponsoda (1996).

Se dispone en el mercado de varios programas que permiten construir tests adaptativos informatizados integrando todas las fases del proceso: calibración del banco de ítems, selección de la estrategia adaptativa, presentación de los ítems, puntuación de los sujetos, etc. El más difundido es el MicroCAT (Assessment Systems Corporation, 1994).

6 - Perspectivas

En los apartados precedentes se han descrito las características generales de los tests informatizados y algunas de sus variantes más extendidas. Sin embargo, la gran actividad investigadora que se está produciendo en este ámbito y el creciente desarrollo tecnológico de los ordenadores permiten augurar la aparición

de nuevos modelos y una fuerte expansión de esta metodología en los próximos años. Aquí presentaremos brevemente las principales líneas de desarrollo futuro: los tests autoadaptados, el desarrollo de nuevos modelos TRI y el diseño de sistemas expertos para la generación automática de ítems.

Tests Autoadaptados

Algunos autores han indicado que los TAI pueden incrementar la ansiedad del sujeto produciendo interferencias cognitivas que podrían influenciar los tiempos de respuesta y decrementar la precisión (Schoonman, 1989). En los últimos años se esta investigando en un tipo de tests que pretenden aminorar este problema: Los Tests Autoadaptados Informatizados (TADI). En comparación con los Tests Adaptativos Informatizados (TAI), los TADI pretenden reducir el nivel de ansiedad con que se afronta la tarea, sin por ello reducir la precision de las estimaciones ni alterar los niveles estimados de habilidad (Rocklin, 1994; Rocklin & O'Donnell, 1987; Wise, Plake, Johnson, & Ross, 1992). En estos tests se agrupan los ítems en subtests de dificultad homogenea (facil, dificultad media, dificil, etc). Al inicio de la prueba, se le pregunta al sujeto en qué nivel de competencia se sitúa (especificada en categorías desde *muy alta* a *muy baja*) a fin de presentarle ítems adecuados a su competencia o se le pide que indique la dificultad del ítem al que ha de responder (desde *muy facil* a *muy dificil*). Una vez emitida la respuesta, se presenta un feedback sobre la precisión (acierto/error) y se le pide que elija la dificultad del ítem siguiente. El sistema selecciona el ítem más informativo de la categoría correspondiente de dificultad. El proceso continúa hasta obtener una precisión predeterminada. El lector podría sorprenderse por este sistema de evaluación, pero recuérdese que uno de los principios de la TRI es que la puntuación del sujeto es invariante a la muestra de ítems a la que responda. Los datos disponibles ponen de relieve que los TADI reducen los niveles de ansiedad/estado (Rocklin, 1994; Rocklin, O'Donnell, & Holst, 1995). Wise (1994) considera que la reducción puede explicarse desde la teoría del control percibido, según la cual la ansiedad disminuye cuando se considera que se controla el proceso de evaluación. No obstante, aún existen problemas por resolver. Por ejemplo, para obtener estimaciones precisas en el nivel de habilidad, es necesario diseñar la prueba para evitar las nocivas influencias de que un sujeto elija un nivel de dificultad inadecuado. Tampoco se ha investigado suficientemente el efecto del feedback en la violación del supuesto de independencia local. En cualquier caso, los TADI son una vía metodológica muy interesante puesto que incrementan la motivación de los sujetos y las actitudes positivas hacia el proceso de evaluación (Renom, 1993).

Nuevos modelos TRI

Los TAI que se usan en la actualidad se basan fundamentalmente en los modelos logísticos unidimensionales de uno, dos y tres parámetros aplicados a respuestas dicotómicas (típicas de los tests de aptitudes y rendimiento). Sin embargo, muchos tests de actitudes y personalidad utilizan respuestas policotómicas. En los últimos años, ha comenzado la aplicación de varios modelos dirigidos a este tipo de respuestas, lo cuál permitirá expandir la aplicación de la TRI y de los tests

informatizados más allá del ámbito de las pruebas de ejecución máxima. El modelo de respuesta gradual de Samejima es el apropiado para formatos de respuesta tipo Likert. En este modelo se calcula una curva característica para categoría de respuesta que propociona la probabilidad de que sujetos con distinto nivel en el rasgo elijan la categoría. Cuando las categoría de respuesta a un ítem no son ordinales sino nominales (por ejemplo, ¿Que prefiere hacer en sus ratos de ocio?: leer, salir con los amigos, ver la TV...), podría ser util el modelo de respuesta nominal de Bock. Sin duda el supuesto de unidimensionalidad es el más duro de los modelos más extendidos, de forma que no son aplicables a muchos fenómenos psicológicos. En los últimos años se esta investigando en modelos multidimensionales (Embretson, 1997; Fischer & Seliger, 1997; Reckase, 1997). Una excelente recopilación de los nuevos modelos y del software apropiado puede encontrarse en Van der Linden y Hambleton (1996).

Generación automática de ítems

La técnica denominada *generación automática de ítems* tiene sus bases en el diseño de tests a partir del procesamiento de la información, que es una metodología derivada de las nuevas corrientes de interdisciplinariedad entre la psicometría y la psicologia cognitiva (Snow & Lohman, 1989). Desde la perspectiva cognitiva, los procedimientos psicométricos tradicionales para construir tests han sido considerados inadecuados para describir los constructos psicológicos porque se basa en el análisis de los productos (los resultados de un test) y no del proceso de ejecución (las operaciones mentales que llevan a cabo los sujetos para resolver las tareas). Desde el enfoque cognitivo, los ítems de los tests son considerados como tareas susceptibles de ser analizadas en el laboratorio para contrastar modelos alternativos de los procesos que median entre su presentación y la emisión de la respuesta. La formulación y contrastación de modelos aporta algunas informaciones de sumo interés para la construcción de ítems. Las más destacables son la determinación de los procesos más influyentes en la ejecución de los ítems y la identificación de las características o condiciones de los ítems que suscitan el funcionamiento de un proceso mental. El primer aspecto permite dotar de significación teórica a las puntuaciones de los ítems. Es decir, fundamenta la validación de constructo. El segundo permite establecer las reglas para generar los ítems que son apropiados para medir el constructo. Los modelos, que pueden variar en simplicidad (número de procesos, tipos de secuenciamiento, etc.), han sido formulados en múltiples ámbitos sustantivos. Ejemplos notables son los trabajos de Frederiksen en Comprensión Verbal (1981, 1982), Sternberg en Razonamiento (1977), y Embretson (1993), Mumaw, Pellegrino y Glaser (1980) y Pellegrino y Kail (1982) en Aptitud Espacial.

Bejar (1993) ha demostrado que conocer la *carga cognoscitiva* y los parámetros TRI asociados a las facetas de los ítems (las propiedades psicométricas y los procesos cognoscitivos asociados a las condiciones del material estimular) puede evitar la necesidad de construir un gran banco de ítems. Sólo es necesario diseñar una tarea inicial de la que se generan variantes mediante unas reglas de transformación automática. Por el momento, sólo se ha propuesto este procedimiento en el ámbito de los tests de ejecución máxima.

Referencias

Allegiant (1996). *SuperCard.* San Diego, CA: Allegiant Technologies, Inc.

APA (1986). *Guidelines for Computer-Based Tests and Interpretations.* Washington, DC: American Psychological Association.

Assessment Systems Corporation (1994). *User's Manual for MicroCAT. The MicroCAT Testing System.* Assessment Systems Corporation St. Paul, Minnesota.

Backhoff, E., Ibarra, M.A., & Rosas, M. (1997). Evaluación por computadora: una nueva tecnología para la aplicación de exámenes de admisión. *Psicología Contemporánea, 4,* 4-11.

Barbero, M. (1996). Banco de ítems. En J. Muñiz (Ed.). *Psicometría.* Madrid: Univérsitas.

Bartram, D. (1993). Emerging trends in computer-assisted assessment. En H. Schuler, J.L. Farr y M. Smith (Eds.), *Personnel selection and assessment.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Bejar, I.I. (1993). A Generative Approach to Psychological and Educational Measurement. En N. Frederiksen, R.J. Mislevy y I.I. Bejar (Eds.), *Test theory for a new generation of tests* . Hillsdale, N.J.: LEA.

Booth, P. (1989). *An introduction to Human-Computer Interaction.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Brand, N. & Houx, P.J. (1992). MINDS: Towards a computerized test battery for use in health psychological and neuropsichological assessment. *Behavior Research Methods, Instrument & Computers, 24,* 385-389.

Bunderson, C.V., Inouye, D.K., & Olsen, J.B.) (1989). The Four Generations of Computerized Educational Measurement. En N. Frederiksen, R.J. Mislevy e I.I. Bejar (Eds.), *Test Theory for a new generation of tests.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Carroll, J.B. (1987). New Perspectives in the Analysis of Abilities. En R.R. Ronning, J.C. Conoley, J.A. Glover,& J.C. Witt (Eds.), *The Influence of Cognitive Psychology on Testing*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Cole, N.S. & Moss, P.A. (1989). Bias in Test Use. En R.L. Linn (Ed.), *Educational Measurement* (201-219). New York: Mcmillan.

Embretson, S. (1993). Psychometric models for learning and cognitive processes. En N. Frederiksen, R.J. Mislevy y I.I. Bejar (Eds.), *Test theory for a new generation of tests* (125-150). Hillsdale, N.J.: LEA.

Embretson, S.. (1997). Multicomponent response models. En W.J. Van der Linden y R.K. Hambleton (Eds.), *Handbook of modern item response theory.* New York: Springer-Verlag.

Evan, W.M. & Miller, J.R. (1969). Differential effects on response bias of computer versus conventional administration of a social science questionnaire. *Behavior Science, 14,* 216-227.

Fischer, G.H. & Seliger, E. (1997). Multidimensional linear logistic models for change. En W.J. Van der Linden y R.K. Hambleton (Eds.), *Handbook of modern item response theory.* New York: Springer-Verlag.

Fowler, R.D. (1987). Developing a computer-based test interpretation system. En J.N. Butcher (Ed.), *Computerized Psychological Assessment: A practitioner's Guide.* New York: Basic Books.

Frederiksen, J.R. (1981). Sources of process interaction in reading. En A.M. Lesgold y C.A. Perfetti (Eds.). *Interactive processes in reading* (pp 361-386). Hillsdale, N.J.: LEA.

Frederiksen, J.R. (1982). A componential theory of reading skills and their interactions. En R.J. Sternberg (Ed.), *Advances in psychology of human intelligence. Vol. 1.* (125-180). Hillsdale, N.J.: LEA.

Green, B.F. (1990). System Design and Operations. En H. Wainer (Ed.), *Computerized Adaptive Testing: A Primer.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item response theory.* Beverly Hills, CA: Sage.

Hambleton, R.K., Zaal, J.N., & Pieters, J.P.M. (1991). Computerized Adaptive Testing: Theory, Applications, and Standards. En R.K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing.* Boston: Kluwer Academic Publishers.

Held, J.J., O'Neil, H.F., & Hansen, D.N. (1973). Affective reactions toward computer-based intelligence testing. *Journal of Consulting and Clinical Psychology, 40,* 217-222.

Kyllonen, P.C. (1991). Principles for creating a computerized test battery. *Intelligence, 15,* 1-15.

Lohman, D.F. & Ippel, M.J. (1993). Cognitive Diagnosis: From Statistically Based Assessment Toward Theory-Based Assessment. En N. Frederiksen, R.J. Mislevy, & I.I. Bejar (Eds.), *Test Theory for a new generation of tests.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Lopez, J.A., Ato, M., Sanchez, J., & Velandrino, A.P. (1990). Test y diagnóstico psicológico por ordenador. En S. Algarabel & J. Sanmartín (Eds.), *Métodos informáticos aplicados a la psicología.* Madrid: Pirámide.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: LEA.

Lushene, R.E., O'Neil, O.F., & Dunn, T. (1974). Equivalent validity of a completely computerized MMPI. *Journal of Personality Assessment, 38,* 353-361.

Mazzeo, J. & Harvey, A.L. (1988). *The equivalence of scores from automated & conventional versions of educational & psychological tests: A review of the literature* (Research Report No. CBR 87-8, ETS RR 88-21). Princeton, NJ: Educational Testing Service.

Mislevy, R.J. & Bock, R.D. (1984). *BILOG: Maximum likelihood item analysis and test scoring with logistic models.* Mooresville, IN: Scientific Software.

Mumaw, R.J., Pellegrino, J.W., & Glaser, R. (1980). *Some puzzling aspects of spatial ability.* Paper presented at annual meetings of the Psychonomic Society. St. Louis, MO, November.

Muñiz, J. (1996). *Psicometría.* Madrid: Univérsitas.

Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Items.* Madrid: Pirámide.

Muñiz, J. & Hambleton, R.K. (1992). Medio siglo de teoría de respuesta a los ítems. *Anuario de psicología, 52,* 41-66.

Neubauer, A.C. & Malle, B.F. (1997). Questionnaire Reponse Latencies: Implications for Personality Assessment and Self-Schema Theory. *European Journal of Psychological Assessment, 13,* 109-117.

Olea, J. & Ponsoda, V. (1996). Tests adaptativos informatizados. En J. Muñiz (Ed.), *Psicometría.* Madrid: Univérsitas.

Pasquali, L. (1996). *Teoria e métodos de medida em ciências do comportamento.* Brasília: Laboratorio de Pesquisa em Avaliaçao e Medida/Instituto de Psicologia/ UnB:INEP.

Pellegrino, J.W. & Kail, R.V. (1982). Process analyses of spatial aptitude. En R.J. Sternberg (Ed.), *Advances in psychology of human intelligence. Vol. 1.* (311-365). Hillsdale, N.J.: LEA.

Plutchik, R. & Karasu, T.B. (1991). Computers and Psychoterapy: An overview. *Computers in Human Behavior, 7,* 33-44.

Prieto, G. & Delgado, A.R. (1996). Construcción de los ítems. En J. Muñiz (Ed.), *Psicometría.* Madrid: Univérsitas.

Reckase, M.D. (1997). A linear logistic multidimensional model for dichotomous item response data. En W.J. Van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory.* New York: Springer-Verlag.

Renom, J. (1993). *Tests adaptativos computerizados.* Barcelona: PPU.

Rocklin, T.R. (1994). Self-adapted testing. *Applied Measurement in Education, 7,* 3-14

Rocklin, T.R. & O'Donnell, A.M. (1987). Self-adapted testing: A performance improving variant of computerized adaptive testing. *Journal of Educational Psychology, 79,* 315-319.

Rocklin, T.R., O'Donnell, A.M., & Holst, P.M. (1995). Effect and underlying mechanisms of Self-Adapted Testing. *Journal of Educational Psychology, 87,* 103-116.

Rolls, S. & Feltham, R. (1993). Practical and professional issues in computer-based assessment and interpretation. En M. Smith & V. Sutherland (Eds.), *Professional issues in selection and assessment.* Chichester, England: John Wiley & Sons.

Ronning, R.R., Conoley, J.C., Glover, J.A., & Witt, J.C. (Eds.).(1987). *The influence of cognitive psychology on testing.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Schmitt, N., Gilliland, S.W., Landis, R.S., & Devine, D. (1993). Computer-based testing applied to selection of secretarial applicants. *Personnel Psychology*, *46*, 149-165.

Schoonman, W. (1989). *An applied study on computerized adaptive testing.* Amsterdam/Lisse: Swets & Zeitlinger.

Snow, R. E. & Lohman, D.F. (1989). Implications of Cognitive Psychology for Educational Measurement. En R.L. Linn (Ed.), *Educational Measurement.* New York: Macmillan.

Sternberg, R.J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities.* Hillsdale, N.J.: LEA.

Sweetland, R.C. & Keyser, D.J. (1991). *Tests: A comprehensive reference for assessments in psychology, education and business.* Austin, TX: Pro-ed.

SPSS, Inc. (1994). *SPSS windows user's guide.* New York: MacGraw Hill.

Van der Linden, W.J. & Hambleton, R.K. (Eds. - 1996). *Handbook of modern item response theory.* New York: Springer-Verlag.

Wingersky, M.S. & Lord, F.M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement*, *8*, 347-364.

Wise, S.L. (1994). Guest editor's note. *Applied Measurement in Education*, *7*, 1.

Wise, S.L., Barnes, L.B., Harvey, A.L., & Plake, B.S. (1989). Effects of computer anxiety and computer experience on computer-based achievement test performance of college students. *Applied Measurement in Education*, *2*, 235-241.

Yen, W.M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 245-262.

Como Elaborar um Questionário

Hartmut Günther[1]

São três os caminhos principais para compreender o comportamento humano no contexto das ciências sociais empíricas: (1) observar o comportamento que ocorre naturalmente no âmbito real; (2) criar situações artificiais e observar o comportamento ante tarefas definidas para essas situações; (3) perguntar às pessoas sobre o que fazem (fizeram) e pensam (pensaram). Cada uma das três famílias de técnicas para conduzir estudos empíricos - observação, experimento e *survey* - apresenta vantagens e desvantagens distintas (Kish, 1987). Tais vantagens estão ligadas à qualidade e à utilização dos dados obtidos, a serem consideradas pelo pesquisador quando escolher a mais apropriada para seu objetivo de pesquisa. Não obstante as variações dentro de cada uma destas três grandes áreas, podemos afirmar que o ponto forte da observação é o realismo da situação estudada; que o experimento possibilita tanto a randomização de características das pessoas estudadas quanto inferências causais; e que o levantamento de dados por amostragem, ou *survey*, assegura melhor representatividade e permite generalização para uma população mais ampla.

O presente capítulo trata da elaboração de um questionário, instrumento principal para o levantamento de dados por amostragem. Fink & Kosecoff (1985) definem *survey*, termo inglês geralmente traduzido como *levantamento de dados*, como "método para coletar informação de pessoas acerca de suas idéias, sentimentos, planos, crenças, bem como origem social, educacional e financeira" (p. 13). Importante apontar que 'levantamento de dados' traduz, apenas, o termo *survey*. Como dados também são levantados através de observações, de experimentos, de busca em arquivos, além da interação pergunta-resposta, será utilizado o termo *survey* neste capítulo. O segundo ponto a observar é que, embora a qualificação 'por amostragem' seja necessária para que os resultados de um *survey* possam ser generalizados para uma população maior, não entraremos em detalhes sobre a questão, concentrando o capítulo na construção de um questionário[2].

O instrumento utilizado no *survey*, o *questionário*, pode ser definido como "um conjunto de perguntas sobre um determinado tópico que não testa a habi-

1 Instituto de Psicologia, Universidade de Brasília.
2 Vide nota de rodapé 3.

Fischer, G.H. & Seliger, E. (1997). Multidimensional linear logistic models for change. En W.J. Van der Linden y R.K. Hambleton (Eds.), *Handbook of modern item response theory.* New York: Springer-Verlag.

Fowler, R.D. (1987). Developing a computer-based test interpretation system. En J.N. Butcher (Ed.), *Computerized Psychological Assessment: A practitioner's Guide.* New York: Basic Books.

Frederiksen, J.R. (1981). Sources of process interaction in reading. En A.M. Lesgold y C.A. Perfetti (Eds.). *Interactive processes in reading* (pp 361-386). Hillsdale, N.J.: LEA.

Frederiksen, J.R. (1982). A componential theory of reading skills and their interactions. En R.J. Sternberg (Ed.), *Advances in psychology of human intelligence. Vol. 1.* (125-180). Hillsdale, N.J.: LEA.

Green, B.F. (1990). System Design and Operations. En H. Wainer (Ed.), *Computerized Adaptive Testing: A Primer.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item response theory.* Beverly Hills, CA: Sage.

Hambleton, R.K., Zaal, J.N., & Pieters, J.P.M. (1991). Computerized Adaptive Testing: Theory, Applications, and Standards. En R.K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing.* Boston: Kluwer Academic Publishers.

Held, J.J., O'Neil, H.F., & Hansen, D.N. (1973). Affective reactions toward computer-based intelligence testing. *Journal of Consulting and Clinical Psychology, 40,* 217-222.

Kyllonen, P.C. (1991). Principles for creating a computerized test battery. *Intelligence, 15,* 1-15.

Lohman, D.F. & Ippel, M.J. (1993). Cognitive Diagnosis: From Statistically Based Assessment Toward Theory-Based Assessment. En N. Frederiksen, R.J. Mislevy, & I.I. Bejar (Eds.), *Test Theory for a new generation of tests.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Lopez, J.A., Ato, M., Sanchez, J., & Velandrino, A.P. (1990). Test y diagnóstico psicológico por ordenador. En S. Algarabel & J. Sanmartín (Eds.), *Métodos informáticos aplicados a la psicología.* Madrid: Pirámide.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: LEA.

Lushene, R.E., O'Neil, O.F., & Dunn, T. (1974). Equivalent validity of a completely computerized MMPI. *Journal of Personality Assessment, 38,* 353-361.

Mazzeo, J. & Harvey, A.L. (1988). *The equivalence of scores from automated & conventional versions of educational & psychological tests: A review of the literature* (Research Report No. CBR 87-8, ETS RR 88-21). Princeton, NJ: Educational Testing Service.

Mislevy, R.J. & Bock, R.D. (1984). *BILOG: Maximum likelihood item analysis and test scoring with logistic models.* Mooresville, IN: Scientific Software.

Mumaw, R.J., Pellegrino, J.W., & Glaser, R. (1980). *Some puzzling aspects of spatial ability.* Paper presented at annual meetings of the Psychonomic Society. St. Louis, MO, November.

Muñiz, J. (1996). *Psicometría.* Madrid: Univérsitas.

Muñiz, J. (1997). *Introducción a la Teoría de Respuesta a los Items.* Madrid: Pirámide.

Muñiz, J. & Hambleton, R.K. (1992). Medio siglo de teoría de respuesta a los ítems. *Anuario de psicología, 52,* 41-66.

Neubauer, A.C. & Malle, B.F. (1997). Questionnaire Reponse Latencies: Implications for Personality Assessment and Self-Schema Theory. *European Journal of Psychological Assessment, 13,* 109-117.

Olea, J. & Ponsoda, V. (1996). Tests adaptativos informatizados. En J. Muñiz (Ed.), *Psicometría.* Madrid: Univérsitas.

Pasquali, L. (1996). *Teoria e métodos de medida em ciências do comportamento.* Brasília: Laboratorio de Pesquisa em Avaliaçao e Medida/Instituto de Psicologia/ UnB:INEP.

Pellegrino, J.W. & Kail, R.V. (1982). Process analyses of spatial aptitude. En R.J. Sternberg (Ed.), *Advances in psychology of human intelligence. Vol. 1.* (311-365). Hillsdale, N.J.: LEA.

Plutchik, R. & Karasu, T.B. (1991). Computers and Psychoterapy: An overview. *Computers in Human Behavior, 7,* 33-44.

Prieto, G. & Delgado, A.R. (1996). Construcción de los ítems. En J. Muñiz (Ed.), *Psicometría.* Madrid: Univérsitas.

Reckase, M.D. (1997). A linear logistic multidimensional model for dichotomous item response data. En W.J. Van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory.* New York: Springer-Verlag.

Renom, J. (1993). *Tests adaptativos computerizados.* Barcelona: PPU.

Rocklin, T.R. (1994). Self-adapted testing. *Applied Measurement in Education, 7,* 3-14

Rocklin, T.R. & O'Donnell, A.M. (1987). Self-adapted testing: A performance improving variant of computerized adaptive testing. *Journal of Educational Psychology, 79,* 315-319.

Rocklin, T.R., O'Donnell, A.M., & Holst, P.M. (1995). Effect and underlying mechanisms of Self-Adapted Testing. *Journal of Educational Psychology, 87,* 103-116.

Rolls, S. & Feltham, R. (1993). Practical and professional issues in computer-based assessment and interpretation. En M. Smith & V. Sutherland (Eds.), *Professional issues in selection and assessment.* Chichester, England: John Wiley & Sons.

Ronning, R.R., Conoley, J.C., Glover, J.A., & Witt, J.C. (Eds.).(1987). *The influence of cognitive psychology on testing.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Schmitt, N., Gilliland, S.W., Landis, R.S., & Devine, D. (1993). Computer-based testing applied to selection of secretarial applicants. *Personnel Psychology*, *46*, 149-165.

Schoonman, W. (1989). *An applied study on computerized adaptive testing.* Amsterdam/Lisse: Swets & Zeitlinger.

Snow, R. E. & Lohman, D.F. (1989). Implications of Cognitive Psychology for Educational Measurement. En R.L. Linn (Ed.), *Educational Measurement.* New York: Macmillan.

Sternberg, R.J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities.* Hillsdale, N.J.: LEA.

Sweetland, R.C. & Keyser, D.J. (1991). *Tests: A comprehensive reference for assessments in psychology, education and business.* Austin, TX: Pro-ed.

SPSS, Inc. (1994). *SPSS windows user's guide.* New York: MacGraw Hill.

Van der Linden, W.J. & Hambleton, R.K. (Eds. - 1996). *Handbook of modern item response theory.* New York: Springer-Verlag.

Wingersky, M.S. & Lord, F.M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement*, *8*, 347-364.

Wise, S.L. (1994). Guest editor's note. *Applied Measurement in Education*, *7*, 1.

Wise, S.L., Barnes, L.B., Harvey, A.L., & Plake, B.S. (1989). Effects of computer anxiety and computer experience on computer-based achievement test performance of college students. *Applied Measurement in Education*, *2*, 235-241.

Yen, W.M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 245-262.

Como Elaborar um Questionário

Hartmut Günther[1]

São três os caminhos principais para compreender o comportamento humano no contexto das ciências sociais empíricas: (1) observar o comportamento que ocorre naturalmente no âmbito real; (2) criar situações artificiais e observar o comportamento ante tarefas definidas para essas situações; (3) perguntar às pessoas sobre o que fazem (fizeram) e pensam (pensaram). Cada uma das três famílias de técnicas para conduzir estudos empíricos - observação, experimento e *survey* - apresenta vantagens e desvantagens distintas (Kish, 1987). Tais vantagens estão ligadas à qualidade e à utilização dos dados obtidos, a serem consideradas pelo pesquisador quando escolher a mais apropriada para seu objetivo de pesquisa. Não obstante as variações dentro de cada uma destas três grandes áreas, podemos afirmar que o ponto forte da observação é o realismo da situação estudada; que o experimento possibilita tanto a randomização de características das pessoas estudadas quanto inferências causais; e que o levantamento de dados por amostragem, ou *survey*, assegura melhor representatividade e permite generalização para uma população mais ampla.

O presente capítulo trata da elaboração de um questionário, instrumento principal para o levantamento de dados por amostragem. Fink & Kosecoff (1985) definem *survey*, termo inglês geralmente traduzido como *levantamento de dados*, como "método para coletar informação de pessoas acerca de suas idéias, sentimentos, planos, crenças, bem como origem social, educacional e financeira" (p. 13). Importante apontar que 'levantamento de dados' traduz, apenas, o termo *survey*. Como dados também são levantados através de observações, de experimentos, de busca em arquivos, além da interação pergunta-resposta, será utilizado o termo *survey* neste capítulo. O segundo ponto a observar é que, embora a qualificação 'por amostragem' seja necessária para que os resultados de um *survey* possam ser generalizados para uma população maior, não entraremos em detalhes sobre a questão, concentrando o capítulo na construção de um questionário[2].

O instrumento utilizado no *survey*, o *questionário*, pode ser definido como "um conjunto de perguntas sobre um determinado tópico que não testa a habi-

1 Instituto de Psicologia, Universidade de Brasília.
2 Vide nota de rodapé 3.

lidade do respondente, mas mede sua opinião, seus interesses, aspectos de personalidade e informação biográfica" (Yaremko, Harari, Harrison, & Lynn., 1986, p. 186). Observa-se que a maneira de apresentar o conjunto de perguntas não faz parte da definição. O questionário pode ser administrado em interação pessoal: em forma de entrevista individual ou por telefone; e pode ser auto-aplicado: após envio por correio ou em grupos. Nas definições de *survey* - e questionário - está implícita sua aplicabilidade às mais diversas áreas das ciências sociais.

Neste capítulo trataremos do desenvolvimento de um instrumento para *survey* em cinco seções. A princípio, instrumento e questionário são considerados sinônimos. A primeira seção lidará com as *bases conceituais e populacionais* de um questionário. A segunda tratará do *contexto social da aplicação do instrumento*. A seguir apresenta-se a *estrutura lógica do instrumento;* e na quarta seção, os *elementos do instrumento*, i.é, questões e itens. Na quinta seção apontam-se *diferenças nos instrumentos*, conforme a maneira de sua aplicação: entrevista individual, pelo telefone, por correio convencional ou eletrônico, ou em grupos.

1 - Base Conceitual e Populacional do Questionário

Na elaboração de um questionário para um *survey*, deve-se partir da seguinte reflexão: qual o objetivo da pesquisa em termos dos conceitos a serem pesquisados e da população alvo? Utilizando-se como ponto de partida as considerações de Schuman e Kalton (1985), sumarizadas na Figura 10-1, verifica-se que os objetivos de uma pesquisa levam necessariamente à relação conceito / item e à relação população alvo / amostra. Os dois binômios são correspondentes: *item* e *amostra* constituem a parte prática dos termos abstratos *conceito* e *população*, respectivamente. No desenvolvimento precisam ser tratados paralelamente, i.é, ao determinar os itens em função dos conceitos subjacentes há que levar em conta o binômio população alvo / amostra, da mesma maneira que a determinação da amostra a partir de uma população alvo exige consideração do binômio conceito / item.

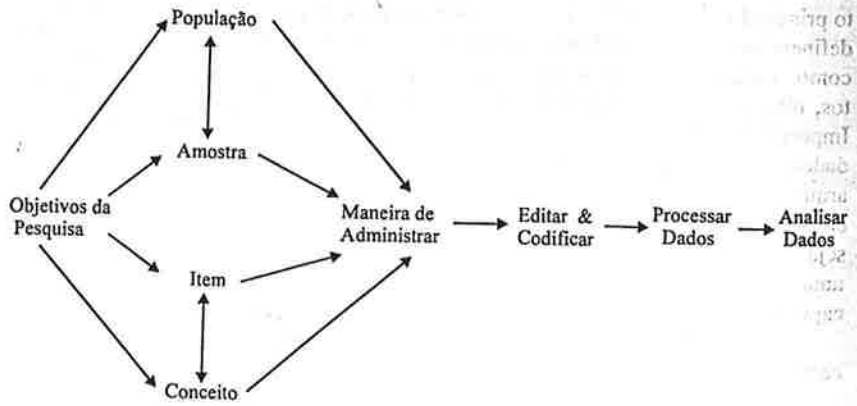


Fig. 10-1. Estágios principais de um Survey
Fonte: Schuman & Kalton (1985)

Imaginemos uma pesquisa tipo *survey* com o *objetivo* de conhecer opções de lazer (*conceito*) entre jovens (*população alvo*) de uma cidade. Lembramos, inicialmente, que os objetivos podem ser muito diferentes. Por exemplo: (a) avaliar as opções existentes (esporte, teatros, cinema, bares, clubes, etc.); (b) levantar a necessidade de opções novas e/ou adicionais; ou (c) estudar entre os jovens o bem-estar psicológico relacionado às opções de lazer disponíveis.

1.1 - Conceito – Itens

O *objetivo* do estudo determinará os *conceitos* a serem investigados num dado *survey*. No presente exemplo, e conforme a definição do objetivo, deve-se diferenciar conceitualmente (a) avaliação de algo existente e (b) levantamento de necessidades de algo inexistente, além de (c) distinguir entre existência ou falta de algum objeto externo ao indivíduo e de um estado de espírito interno. Embora objetivos e conceitos não sejam mutuamente excludentes, podendo até ser tratados numa mesma pesquisa, não faltam exemplos de confusão entre si.

Num segundo momento, o objetivo do estudo determina as perguntas concretas a serem apresentadas (i.é, os itens), além de existir uma relação recíproca entre *conceitos* e *itens*. Dependendo dos *conceitos* a serem pesquisados, o conteúdo das perguntas ou *itens* varia. Igualmente, a possibilidade de fazer certas perguntas (mais facilmente) a determinadas pessoas (amostra) faz com que um ou outro *conceito* possa ser explorado numa dada pesquisa. Finalmente, os conceitos subjacentes e, especialmente, o conteúdo dos itens, determinam o instrumento e a maneira da sua apresentação.

1.2 – População alvo – Amostra

Não é apenas o objetivo da pesquisa que determina a forma do instrumento via conceitos e itens. Dependendo do objetivo, a *população alvo* pode ser constituída, no nosso exemplo, por 'todos' os jovens da cidade. Ou por determinados subgrupos, como crianças, meninos, meninas ou adolescentes. Dependendo do objetivo da pesquisa e da população alvo (especialmente do seu tamanho e da sua acessibilidade), definir-se-ão diferentes tipos de *amostra*. Se a população alvo for constituída por 'todos' os jovens da cidade, é necessário que os subgrupos relevantes para o estudo sejam adequadamente representados na amostra. No caso da população ser constituída por um determinado grupo - por exemplo, meninas adolescentes - o número e a maneira de selecionar os participantes do estudo dentro do grupo sofreria modificação[3].

Embutida na seleção da população alvo e da amostra estão as características da amostra: nível educacional, idade, gênero. Apesar do fato de que a determinação destas características (p. ex., percentual de meninas alfabetizadas com 12

3 A questão da relação entre a *população-alvo* e a *amostra* dos participantes é assunto de livros especializados sobre amostragem. O tema não será tratado neste capítulo, uma vez que uma introdução detalhada é fornecida por Kish (1965). A maioria dos textos de metodologia da pesquisa social inclui capítulos sobre amostragem (p. ex., Aday, 1989; Babbie, 1992; Frankel, 1983; Judd, Smith, & Kidder, 1992; Sudman, 1983). Pode-se sugerir: Henry (1990, 1998) ou Williams (1978), dentre muitas outras fontes.

anos de idade) pode ser, por si só, objetivo de um *survey*, estimativas - embora vagas - sobre as características são necessárias ao se tomarem decisões sobre a maneira de administrar um instrumento (p. ex., uso de instrumento auto-aplicado implica em população alvo alfabetizada).

1.3 - Reciprocidade entre Conceito e População alvo

Sumarizando, o objetivo de uma pesquisa determina a forma do instrumento e a maneira da sua aplicação por meio de *conceitos* e *itens* e da *população alvo* idealizada e da *amostra*. Verifica-se a seguinte interdependência entre a elaboração de um instrumento e a estratégia de sua aplicação: (1) o grau de complexidade dos conceitos determina número de itens e forma de apresentação deles; (2) existe relação recíproca entre as características da população alvo e a complexidade dos conceitos a serem investigados. Ambos determinam a maneira de transformação dos conceitos em itens (perguntas) e sua administração; (3) o tamanho da amostra influencia a maneira de administrar o instrumento em termos de entrevista *vs.* questionário e em termos de tamanho.

Além do mais, o tamanho da amostra é determinado pelos recursos (tempo, dinheiro e recursos humanos) disponíveis. Esta disponibilidade de recursos influencia o planejamento da administração do instrumento, bem como a codificação das respostas, seu processamento e, eventualmente, as possíveis análises. Considerem-se os exemplos seguintes.

Uma pesquisa que visa determinar a satisfação com a disponibilidade de opções básicas de lazer como quadra de esporte, clubes, teatros, pode ter como população alvo "jovens com telefone em casa". Caso o pesquisador tenha à disposição recursos para contratar entrevistadores, treiná-los e instalar dez linhas telefônicas e microcomputadores para registro das respostas no ato da entrevista, será possível levantar dados junto a um grande número de pessoas dentro de pouco tempo.

Quando se pretende explorar conceitos como clima social, confiança mútua e solidariedade entre moradores de um bairro periférico, idealizando como população alvo jovens desconfiados, é exigido outro tipo de instrumento, outra maneira de aplicação e administração, resultando em método de codificação, processamento e análise diferentes.

Em suma, embora este capítulo trate da construção do questionário para *surveys* de maneira geral e faça considerações sobre sua aplicação, o leitor deve lembrar que os detalhes do instrumento dependerão da população alvo, do tamanho da amostra, dos conceitos a serem explorados, bem como dos recursos disponíveis para aplicação e processamento do instrumento. Além de considerar a interdependência de população e amostra, de um lado, e conceitos e itens do outro, ao elaborar um questionário o pesquisador deve atentar para os tópicos tratados nas seções: (a) o contexto social da aplicação do instrumento; (b) a estrutura lógica do instrumento na organização de seus elementos; (c) os elementos do instrumento, i.é, questões e itens; (d) diferenças nos instrumentos.

2 - Contexto Social da Aplicação do Instrumento

Diante do fato de que o respondente de um *survey* gasta seu tempo e faz algum esforço mental, uma reflexão básica deve ser: *quem* deseja algo de *quem* numa determinada pesquisa? A disposição do respondente em revelar algo sobre si mesmo, permitindo ao pesquisador obter os dados desejados, varia conforme a situação. Mencionem-se alguns exemplos: confessionário (padre - fiel), interrogatório (policial - suspeito), declaração de renda (receita - contribuinte), seleção e concurso (comissão de admissão - candidato), prova (professor - aluno), aconselhamento (psicólogo - cliente voluntário), manicômio (psiquiatra - cliente mandado), procura de emprego (funcionário de RH - solicitante), pesquisa de opinião e marketing (firma de pesquisa, podendo ou não oferecer brindes - respondente), pesquisa social acadêmica (pesquisador - 'sujeito').

A este trabalho interessam os dois últimos exemplos, caracterizados pelo fato de que o pesquisador não tem poder sobre o respondente e precisa convencê-lo de que vale a pena participar da pesquisa. A seguir consideramos alguns aspectos do contexto social e cultural desta interação pergunta-resposta: (a) o fundo cultural, (b) o *background* do pesquisador, (c) o contexto da pesquisa e (d) o *background* do respondente (Pareek & Rao, 1980, p. 154)[4].

2.1 - Background Cultural

Até onde se aceita ser indagado por um estranho acerca de assuntos pessoais? Quais assuntos são considerados 'públicos'? Quais são 'privados'? As vertentes são nível de reticência e cortesia (senso de obrigação de agradar o outro) e, finalmente, levar tal interação a sério, fornecendo respostas autênticas.

2.2 - Background do Pesquisador

Nesta dimensão entram as considerações: imagem e afiliação do pesquisador, inclusive a imagem da organização à qual o pesquisador é afiliado; a distância social e cultural pesquisador / organização e respondente; relevância do assunto para o respondente; viés do pesquisador / organização.

2.3 - Contexto da Pesquisa

Além do ambiente físico e social no qual a pesquisa será conduzida (p. ex., na casa do respondente, num local público como shopping, no local de trabalho do pesquisador ou do respondente), relevância e sensitividade temática terão notável influência sobre a disposição do respondente de participar da pesquisa. Acrescenta-se relevância cultural e aspectos de desejabilidade do tema (i.é, à época de Natal será difícil realizar uma pesquisa autêntica sobre a desejabilidade em dar esmolas a crianças pobres). O potencial do instrumento de aprofundar, com tamanho e estrutura adequados, é importante para a consecução de dados válidos.

---

4 Vide também Günther, Brito & Silva (1989) para uma discussão mais ampla deste assunto.

2.4 - Background do Respondente

No que diz respeito às características do respondente, há que considerar inicialmente a distância entre opinião pública e particular. Estamos interessados na opinião do respondente, não na opinião de outros. Relacionada a este aspecto encontra-se a situação do respondente que se considera capaz de opinar sobre qualquer assunto. Embora uma experiência anterior como participante em pesquisa possa ser desejável, a experiência não é recomendável quando se pesquisa um mesmo assunto; quando a participação for recente, não se deve tentar pesquisar o respondente se ele mostrar-se desinteressado.

Ao desenvolver o instrumento, convém lembrar essas dimensões preferencialmente antes de iniciar o processo de planejamento. Apesar de as considerações terem sido levantadas no contexto da pesquisa transcultural, com intuito de ajudar o pesquisador a desenvolver um instrumento a ser aplicado numa cultura que não é a sua, não há razão para supor que o pesquisador conheça sua própria cultura a ponto de se comportar intuitivamente correto diante dos respondentes pretendidos.

3 - Estrutura Lógica do Instrumento

Consideramos, inicialmente, as razões que levam uma pessoa a responder a um instrumento de pesquisa. Falando de *survey* via correio, Dillman (1978) afirma que "o processo de mandar um questionário a respondentes em potencial, conseguir que completem e devolvam o questionário de maneira honesta pode ser visto como caso especial de 'troca social'" (p. 12). Aplicando a teoria de troca social a *survey*, Dillman chega à seguinte conclusão, "Assim, há três coisas que precisam ser feitas para maximizar a resposta a *survey*: minimize o custo para o respondente, maximize as recompensas para fazê-lo e estabeleça confiança de que a recompensa será concedida" (p. 12). Traduzida em detalhes operacionais, o autor aponta as ações que um pesquisador poderia fazer num *survey* (Dillman, 1978, p. 18):

1) Recompensar o respondente: a) demonstrando consideração; b) oferecendo apreciação verbal usando uma abordagem consultiva; c) apoiando seus valores; d) oferecendo recompensas concretas; e) tornando o instrumento interessante;
2) Reduzir o custo de responder: a) fazendo com que a tarefa pareça breve; b) reduzindo esforços físico e mental requeridos; c) eliminando a possibilidade de embaraço; d) eliminando qualquer implicação de subordinação; e) eliminando qualquer custo financeiro imediato;
3) Estabelecer confiança: a) oferecendo um sinal de apreciação antecipadamente; b) identificando-se com uma instituição conhecida e legitimada; c) aproveitando outros relacionamentos de troca.

Consideramos que as recomendações se aplicam a qualquer tipo de *survey*, não apenas aos enviados por correio. Pode-se observar a estrutura de um instrumento de *survey* de maneira mais ampla. Numa afirmação clássica, Bingham e Moore (1934) definem entrevista como "conversa com um objetivo". Da mesma maneira que qualquer interação social consiste em um cumprimento, na interação em si e em uma despedida, o instrumento que estrutura a interação entre pesquisador e respondente num *survey* deve refletir as três fases. No cumprimento (a introdução)

reconhece-se o outro e estabelece-se o nível de confiança apropriado e necessário. Segue-se a transação social em si, a interação pergunta - resposta. Na despedida reforça-se qualquer sinalização de benefícios (futuros) já demonstrada. Estes segmentos correspondem, de maneira geral, aos pontos 3, 2 e 1 de Dillman, respectivamente.

3.1 - Introdução: Estabelecer Confiança

A primeira tarefa é estabelecer contato com o respondente em potencial e assegurar sua cooperação. Para estabelecer confiança, o pesquisador/entrevistador precisa apresentar-se e indicar com e para quem trabalha. A seguir, precisa capturar o interesse do respondente pelo tema, porque o tema é importante, especialmente para o respondente. Nada melhor para expressar apreciação do que ressaltar o quanto opiniões e experiências do respondente são importantes.

Do ponto de vista prático, o primeiro passo neste processo é uma boa apresentação do instrumento e/ou da pessoa que o administra. Caso o questionário seja remetido pelo correio, irá acompanhado de uma carta de apresentação da qual constará a informação sobre quem está 'por trás' da pesquisa, para que serve. Em se tratando de entrevista, o entrevistador pode explicar quem ele é, para quem trabalha, identificar-se (p. ex., com crachá) ou entregar alguma carta ao *candidato* a respondente. Como os primeiros momentos decidem sobre a disposição do respondente em cooperar, é aí que qualidade e quantidade de informação sobre a pesquisa precisam se concentrar.

No caso de instrumento auto-aplicado (p. ex., enviado via correio), a introdução não somente precisa ser persuasiva, mas deve conter toda a informação necessária para poder agir da maneira esperada pelo pesquisador. Embora se devam incluir indicações claras de como entrar em contato com o responsável pela pesquisa, caso existam dúvidas, o esforço para pedir instruções adicionais pode fazer com que a maioria dos potenciais respondentes ignore o instrumento, em vez de se informar com o pesquisador.

No caso da aplicação pessoal de um instrumento, o entrevistador, apropriadamente treinado, tem oportunidade de explicitar e tirar dúvidas sobre: (a) quem é responsável pela pesquisa, (b) quais os objetivos e (c) o que é que o respondente deve fazer. A ênfase está no 'apropriadamente treinado', porque existe um contínuo desde começar a esclarecer dúvidas, a tentar persuadir pessoas relutantes em responder, até insistir 'no conteúdo da resposta'. Idealmente, as instruções estandardizadas são tão claras que não é necessário pedir esclarecimentos adicionais ao entrevistador. Mas na medida em que boa parte da recompensa que um pesquisador pode oferecer consiste, justamente, na oportunidade de interação social (veja Günther, Brito e Silva, 1989), o entrevistador precisa de treinamento para fornecer informações sem enviesar as eventuais respostas do respondente.

Nas duas situações, instrumento auto-aplicado e interação pessoal, é vantajoso manter contato prévio com os membros da amostra. 'Aviso prévio' pode variar, conforme recursos disponíveis e abrangência da pesquisa, de uma carta até uma campanha de *outdoors* ou anúncios na televisão. Tais avisos da chegada de entrevistadores, explicitando os objetivos da pesquisa, facilitam a recepção do

entrevistador (Solórzano, 1991) e aumentam a disposição para responder a um questionário enviado pelo correio (Gouveia & Günther, 1995).

3.2 - Interação Pergunta-Resposta: Reduzindo o Custo para Responder

Identificação do pesquisador e legitimação dos objetivos da pesquisa são passos que permitem que se *comece* com o *survey*. Como o respondente pode desistir da pesquisa a qualquer momento, continua a necessidade de convencê-lo, de manter seu interesse a cada etapa da interação. Evitar que desista no meio do processo depende da forma e do conteúdo do instrumento.

Foram citados os pontos indicados por Dillmann para reduzir o custo de responder a um *survey*: a) fazendo com que a tarefa pareça breve; b) reduzindo o esforço físico e mental requerido; c) eliminando a possibilidade de embaraço; d) eliminando qualquer implicação de subordinação; e) eliminando qualquer custo financeiro. Rodeghier (1996, p. 9) sumariza regras gerais como a) manter a tarefa do respondente o mais fácil possível, (b) manter o nível de interesse do respondente o mais alto possível e (c) manter o nível de atenção do respondente o mais alto possível.

Os dois primeiros pontos de Dillman e o primeiro de Rodeghier dizem respeito, explicitamente, à estruturação do instrumento. São conseqüência direta da constatação de que, no contexto de pesquisa social aqui considerado, é o pesquisador que deseja algo do respondente. Assim, além de refletir boas maneiras, considerar as necessidades do respondente mostra bom senso por parte do pesquisador. O assunto será tratado na seção 'Estrutura e Seqüência'. Os demais pontos de Rodeghier, manter interesse e atenção do respondente, também serão abordados.

Quanto à possibilidade de embaraço, tratar-se-á deste tema no contexto de perguntas sensíveis a seguir. A questão da subordinação diz respeito ao direito do respondente de poder suspender a qualquer momento sua participação na pesquisa, se desejar. Jamais a participação deve ser obrigatória ou condicionada à concessão de algum benefício (vide a parte *recompensa*).

Quanto a custos financeiros, Dillmann refere-se a selos no caso de uma pesquisa via correio, mas as considerações podem estender-se a transporte para o local da entrevista ou renda perdida pelo entrevistado durante o tempo da entrevista (vide a parte *recompensa*).

3.3 - Despedida: Reforçar Benefícios da Pesquisa

O mínimo de cortesia na despedida consiste em um agradecimento pela 'valiosa colaboração' do respondente, seja de maneira verbalizada após uma entrevista, seja de maneira escrita ao fim do questionário. No que se refere a benefícios tangíveis e imediatos, o maior beneficiado de uma pesquisa é o pesquisador, não o respondente, embora não signifique que a participação em pesquisa não implique benefício para o respondente. Sentir-se importante por ter sua opinião valorizada ou por poder falar e ser ouvido são motivos fortes para que muitas pessoas procurem participar em *surveys*. Salientar a importância da opinião daquele candidato a respondente é condizente para com esses sentimentos.

A maioria dos participantes de *survey* sabe que o pesquisador é o maior beneficiado, que pouco há para o respondente além da satisfação de ter sido ouvido. Por esta razão, não se deve fazer promessas irreais como "Sua participação nesta pesquisa é importante, uma vez que suas respostas resultarão na melhoria da sua vida". A afirmação é irresponsável e antiética, além de ser percebida como engodo pelo respondente, minando a credibilidade das pesquisas.

Comunicar resultados e/ou facilitar o acesso a eles é forma importante de recompensar os respondentes. Se a 'conversa com um objetivo' foi suficientemente interessante para que o participante mantivesse o nível de atenção, os resultados - apresentados em linguagem não acadêmica - também serão. Além do mais, na hipótese de os resultados provocarem reflexão ou conscientização entre os respondentes sobre o tema da pesquisa, os resultados conterão uma semente para possível melhoria da vida dos respondentes, numa dada temática.

3.4 - Recompensas e Incentivos Financeiros

Por serem mais concretos, incentivos/recompensas financeiras e brindes merecem maior reflexão. Aplica-se a estes, mais explicitamente do que aos demais incentivos, a norma de ética 6.14(b) da American Psychological Association (APA): "Psicólogos não oferecem incentivos financeiros excessivos ou impróprios ou outros incentivos para obter participantes para pesquisa, especialmente quando isto possa obrigar participação" (APA, 1992, p. 1609). Esclarecendo, a questão não é 'se' mas 'quanto' é possível pagar aos participantes de uma pesquisa. Embora do ponto de vista prático o problema seja muitas vezes resolvido pela limitação de recursos do pesquisador, vale lembrar que quanto maior a distância social e/ou financeira entre pesquisador e respondente, maior a possibilidade de criar dependência no incentivo a ponto de o respondente ser privado do direito de suspender sua participação a qualquer momento e dizer (ou fazer) o que julga necessário para continuar na pesquisa (não por acaso se utiliza a palavra *obrigado*). Considerando a norma de ética 6.06(d) "psicólogos tomam medidas razoáveis para implementar proteções apropriadas dos direitos e do bem-estar dos participantes humanos..." (APA, 1992, p. 1608), parece razoável tentar oferecer recompensa no caso do respondente perder uma renda por causa da sua participação na pesquisa, como é o caso de pessoas que trabalham por conta própria, cujo tempo significa dinheiro, sejam profissionais liberais ou crianças de rua (Günther & al., 1989). Em suma, o pesquisador tem de contrabalançar que o respondente não deve perder recursos e assegurar que a compensação não seja excessivamente generosa ou desproporcional, de modo a causar 'dependência' que possa ser explorada. A questão se apresenta justamente com os socialmente mais fracos (p. ex. crianças, idosos e/ou pobres) ou dependentes (crianças, estudantes ou idosos), visto verificar-se relativa ausência de estudos com respondentes socialmente mais poderosos que o pesquisador.

Mangione (1998) enumera uma série de estudos que demonstram que a oferta de incentivos financeiros tende a aumentar a taxa de resposta em *survey* pelo correio. Afirma que "o que é surpreendente nestes resultados de pesquisa não é o fato de que pagamento adiantado de recompensa tem algum impacto, mas que não parece ser necessária uma grande recompensa para obter taxas de resposta maiores"

(p. 409), mencionando-se valores de aproximadamente um dólar norte-americano. Conclui sua discussão afirmando: "A chave da efetividade parece estar em criar um clima no qual o pagamento antecipado seja visto como algo para se sentir bem, em vez de uma técnica manipulativa que constranja o respondente a participar" (p. 410). Aspecto semelhante é acrescentado por Singer, van Hoewyk e Maher (1998), quando argumentam que o pagamento de incentivos a respondentes não cria, necessariamente, expectativas para futuras participações em pesquisa.

3.5 - Estrutura e Seqüência

Como fazer a tarefa ser breve e fácil, ou pelo menos não torná-la aborrecedora ou aversiva? Uma estrutura bem pensada contribui significativamente para reduzir o esforço físico e/ou mental do respondente, além de assegurar que todos os temas de interesse do pesquisador sejam tratados numa ordem que sugira uma 'conversa com objetivo', mantendo-se o interesse do respondente em continuar. Antes de mais nada, focalizar-se no objetivo da pesquisa, nas perguntas que o pesquisador quer responder por meio dela. Saber claramente por que está incluindo cada item no instrumento. Saber o que as respostas implicam para o andamento da pesquisa. No estudo piloto, haveria margem para uma 'pescaria', i.é, para incluir itens sobre os quais o pesquisador não tem certeza se vale a pena perguntar. Mas o instrumento final deve conter apenas os itens que serão analisados.

Um primeiro princípio de estruturação é direcionar-se do mais geral para o mais específico; do menos delicado, menos pessoal, para o mais delicado, mais pessoal. Esta ordem se aplica a conjuntos temáticos de itens e a um grupo de itens que tratam de uma temática em comum.

Aplicada à seqüência de conjuntos temáticos de itens, significa que o primeiro conjunto de itens/perguntas deve ser mais geral e menos sensível. Esta parte pode até consistir, especialmente em entrevistas pessoais, de uma 'conversa preliminar': Numa pesquisa hipotética entre jovens de um bairro sobre opções de lazer, realizada em forma de entrevista pessoal, a parte formal da coleta de dados poderia ser precedida de perguntas gerais sobre a situação do respondente na cidade e no bairro:

> **Há quanto tempo mora nesta cidade?**
> [Caso apropriado] Onde morava antes?
> **Em geral, está satisfeito em morar aqui?**

As perguntas iniciais serviriam menos para obter informação do respondente e mais para estabelecer um relacionamento de confiança entre respondente e pesquisador. Deve-se atentar para que essas perguntas terão de ser repetidas de maneira formal adiante, dentro da entrevista, enquadrando-se as duas primeiras entre os itens do conjunto sócio-econômico e a terceira no conjunto de itens sobre satisfação com o bairro.

Após conseguir convencer um respondente em potencial a dar sua atenção pelo argumento de que a pesquisa trata de assunto de interesse *do respondente*, não convém começar a interação por perguntas burocráticas (nome, sexo, idade) e até delicadas (renda familiar). Em outras palavras, se o participante concorda em responder a pesquisa, porque considera a temática interessante, a primeira pergunta (e as seguintes) deve(m) tratar desta temática. Conquistado e mantido o interesse do respondente, podem ser levantadas perguntas não tão obviamente relacionadas à temática inicialmente sugerida. Lembre-se o contexto social da pesquisa social. Contrária à situação de concurso ou procura de emprego, na qual o respondente está na situação de pedinte, podendo ser sujeitado a qualquer tipo de ficha a preencher, a pesquisa social representa uma situação na qual o pesquisador é o pedinte. Como itens pessoais e sócio-econômicos podem ter conteúdos sensíveis como idade, nível educacional (não é fácil admitir baixo nível educacional de si ou da família), renda individual ou familiar, chegando a estilo de vida, preferências sexuais etc., os itens só terão respostas autênticas quando o participante desenvolver certo grau de confiança no responsável pela pesquisa (representado, no caso de interação pessoal, pelo entrevistador). Mais uma vez vale a reflexão: 'É necessária esta pergunta; é necessário este item?'.

Somente o último conjunto de itens trata das características sócio-econômicas do respondente. Um erro mais que comum é começar um instrumento com o levantamento de dados pessoais, às vezes até chamando seção de "Identificação". Em se tratando de pesquisa, não convém identificar o respondente. Pelo contrário. Geralmente há que assegurar que a pesquisa não visa identificar indivíduos, mas que perguntas sócio-demográficas como educação, estado civil, sexo, idade, composição da família, renda, tempo de moradia, etc. servem apenas para caracterizar a amostra. Perguntar o nome no início de uma entrevista pessoal pode facilitar trato interpessoal, mas mesmo sem registrá-lo pode contradizer qualquer afirmação sobre o caráter confidencial da entrevista.

Dependendo do assunto, a regra 'do geral para o específico' pode ser aplicada à seqüência dos itens, dentro de um grupo que trata de uma temática em comum. Ressalva-se que em se tratando de um conjunto que constitui uma escala, os itens devem ser misturados para evitar que dois deles sejam apresentados um após o outro, ao se tratarem aspectos semelhantes.

Um segundo princípio de organização do instrumento é que, na medida apropriada, deve seguir uma ordem lógica. Usando uma hipotética pesquisa sobre moradia, pergunta-se inicialmente sobre a cidade, depois sobre o bairro, a rua e o prédio onde o respondente mora. Além de progredir do geral para o específico, aproxima-se do respondente. Uma pergunta sobre o relacionamento entre moradores da cidade é menos pessoal, menos ameaçadora do que sobre o relacionamento do respondente com seu vizinho. Fazendo perguntas mais pessoais só após haver estabelecido bom nível de confiança, o entrevistador contribui para obter respostas autênticas. Assim, perguntas pessoais sobre o respondente constituiriam o último conjunto:

> Concluindo, gostaríamos de fazer algumas perguntas para melhor caracterizar os respondentes desta pesquisa ...

Um terceiro princípio, implícito no segundo, sugere que itens tratando de uma mesma temática fiquem juntos e recebam uma introdução que ajude o respondente a concentrar-se na temática a ser tratada:

> Inicialmente, gostaria de saber sua opinião sobre as opções de lazer neste bairro ...

4 - Elementos do Instrumento

A parte central do instrumento de *survey* são as perguntas, pelas quais se tenta obter a informação desejada. Esta seção do capítulo é iniciada com considerações gerais sobre como escrever bons itens. Seguem-se as considerações temáticas, os aspectos técnicos e as considerações estatísticas.

4.1 - Considerações Gerais

Fowler (1998, p. 344) define um bom item como aquele que gera respostas fidedignas e válidas. Apresenta cinco características básicas: (a) a pergunta precisa ser compreendida consistentemente; (b) a pergunta precisa ser comunicada consistentemente; (c) as expectativas quanto à resposta adequada precisam ser claras para o respondente; (d) a menos que se esteja verificando conhecimento, os respondentes devem ter toda informação necessária; e (e) os respondentes precisam estar dispostos a responder. Para assegurar tais atributos, cada pergunta deve ser específica, breve, clara, além de escrita em vocabulário apropriado e correto.

- Quanto à especificidade, contrasta-se: "Qual o parque que você gosta?" com "Em que parque você gosta de passear?"
- Quanto à brevidade, contrasta-se: "Você pode enumerar as atividades que realiza no parque e se as mesmas envolvem poucas ou muitas colegas?" com "Quais as atividades no parque?"
- Quanto à clareza, contrasta-se: "De maneira geral, o que você pensa sobre se as atividades planejadas para o clube são importantes?" com "Que grau de influência você tem no planejamento das atividades do clube?"
- Quanto a vocabulário, é preciso pensar no nível educacional dos respondentes. A linguagem não pode ser complexa nem simples demais. O princípio básico é realizar no mínimo um estudo piloto, sem procurar supor algo *a priori*.

Como escrever bons itens? Sudman e Bradburn (1982, p. 14) apresentam três regras gerais: (a) controle o impulso de escrever itens específicos antes de haver refletido completamente sobre as perguntas da pesquisa; (b) anote as perguntas da pesquisa e mantenha-as perto enquanto estiver desenvolvendo o questionário; e (c) cada vez que escrever um item, indague: "Por que quero saber disto?" - Responda em termos que lhe ajudem a responder as perguntas da pesquisa. "Seria interessante saber" não é resposta aceitável.

Linguagem e ambigüidade. Quanto à linguagem usada na formulação dos itens, atenta-se inicialmente para sua compreensão pela população alvo da pesquisa. Abreviações, gírias ou termos regionais devem ser evitados, da mesma maneira que termos especiais ou sofisticados que estejam aquém da compreensão da população alvo. O problema da ambigüidade está relacionado à questão da linguagem. O respondente está entendendo o que o pesquisador está perguntando?

Viés e ênfase. A escolha das palavras pode direcionar as respostas. Quando se pergunta sobre utilização de áreas comuns num bloco de residência, pode-se indagar se deve ser 'proibido', 'não permitido', 'evitado' ou 'impedido'. Assim como o aviso 'Proibido Estacionar' ou 'Pede-se não estacionar' provoca comportamentos diferentes, o número de respondentes que concordam com um item que contém a palavra 'proibir' vs. 'não permitir' varia (veja Schuman & Presser, 1981, cap. 11).

Pergunta aberta versus fechada. Uma escolha importante no desenvolvimento de itens diz respeito a perguntas abertas vs. fechadas. A discussão é extensa (Günther & Lopes, 1990; Schuman & Presser, 1981). Pode-se sumarizar a discussão nos seguintes termos: para uma pesquisa inicial, exploratória, não conhecendo a abrangência ou a variabilidade das possíveis respostas, são necessárias perguntas abertas. Uma vez que se conhecem os tópicos geralmente mencionados pelos respondentes acerca de uma dada temática, especialmente quando existem muitos respondentes e/ou pouco tempo, deve-se usar perguntas fechadas. O argumento de que perguntas abertas dão mais liberdade de expressão ao respondente é uma falácia. Segundo Sommer e Sommer, o uso de perguntas fechadas "mostra freqüentemente mais respeito à opinião das pessoas, deixando-as classificar suas respostas como positivas, negativas ou neutras, em vez do pesquisador fazer isto para eles" (1997, p. 130).

Da mesma maneira que perguntas abertas servem no início da entrevista para estabelecer um clima receptivo entre pesquisador e respondente, servem, no fim do levantamento, para capturar justamente aquelas opiniões não cobertas pelos itens fechados. Além de um 'apanhado final' ao concluir o questionário ou a entrevista, as perguntas abertas podem ser feitas no fim de um conjunto de perguntas, vez que servem para reforçar a essencial percepção do respondente de que o pesquisador tem interesse na opinião *dele*, respondente. Há que lembrar: perguntas abertas, especialmente em questionários auto-aplicados, exigem mais esforço do respondente; aumentando o custo de resposta diminui a probabilidade de completar e devolver o questionário.

Sumarizando, enfatize-se que *sempre* convém realizar um estudo piloto para verificar *se* e *como* as perguntas estão sendo entendidas pela população alvo. Esta regra não tem exceção. Quando se trata com nova população alvo ou novo questionário, deve-se realizar um novo pré-teste.

4.2 - Considerações Temáticas

No início do capítulo citamos a definição de Fink e Kosecoff (1985) de *survey* como um "método para coletar informação de pessoas acerca das suas idéias, sentimentos, planos, crenças, bem como origem social, educacional e financeira" (p. 13). Implícita nesta definição está a distinção entre itens que tratam de conhecimento, de atitudes e opiniões e de informação fatual. Para cada uma das três categorias, podemos diferenciar itens mais ameaçadores.

Grau de ameaça de itens. Fazem-se perguntas a determinadas pessoas na expectativa de que as questões lhes digam respeito, que tenham conhecimento e/

ou atitudes e opiniões sobre o assunto. O potencial de uma pergunta afetar um respondente de maneira ameaçadora está implícito nesta constatação. O assunto pode ser não muito familiar; talvez seja desagradável admitir desconhecimento. O tema pode ser sensível para o respondente; p. ex., comportamentos considerados socialmente inaceitáveis. Ao desenvolver itens é necessário verificar até que ponto determinadas perguntas podem constituir ameaça ao respondente. Caso existam razões para supor que o tema é sensível, precisa-se verificar maneiras de obter a informação sem provocar constrangimento. Problema maior do que perder um respondente irritado por uma pergunta é receber respostas não autênticas, pela razão de o respondente ter algo a esconder ou não saber como responder.

Quanto ao constrangimento provocado por falta de lembrança de comportamento, existem maneiras de ajudar o pesquisador a 'ter sucesso': fazer perguntas específicas ao invés de gerais, além de detalhar contextos temporais e/ou espaciais. Pode ser evitado o constrangimento devido à falta de conhecimento, deixando claro que o conjunto de perguntas não constitui um teste e que é natural às pessoas não ter respostas para todos os itens. Implícito nesta afirmação está: deve-se fazer mais que uma pergunta para assegurar avaliação mais discriminada do nível de conhecimento do que seria possível com apenas uma ou duas.

Quanto a perguntas sobre comportamentos socialmente inaceitáveis ou até ilegais, não entraremos nas questões éticas ou jurídicas, mas lembramos que a tradição não garante ao pesquisador proteção contra eventual obrigação de revelar suas fontes. Assim, o pesquisador deve lembrar-se da norma 1.14 da APA "Psicólogos tomam medidas adequadas para evitar prejuízo para seus pacientes, clientes, participantes de pesquisa, estudantes ou outros com as quais trabalham" (APA, 1992, p. 1601).

Havendo decidido ser justificável fazer perguntas sensíveis, a regra básica é utilizar perguntas abertas - sugestão que implica em entrevista pessoal como modo de interação com o respondente. Contribui para tornar uma pergunta menos ameaçadora e contextualizá-la de maneira que sua importância relativa na entrevista seja reduzida.

Menciona-se, adicionalmente, a técnica de resposta randômica pela qual é possível estimar a proporção de respondentes que mantêm atitudes ou comportamentos socialmente inaceitáveis, sem entretanto poder determinar se um determinado respondente assim se comporta. Para maiores detalhes desta técnica veja, por exemplo, Sudman e Bradburn (1982, cap. 3) ou Zdep, Rhodes, Schwarz , e Kilkenny (1989).

Itens para avaliar conhecimento. Embora não seja função da pesquisa social testar habilidades ou conhecimentos no sentido escolar, sua verificação dentro de uma pesquisa é importante. Perguntas de conhecimento importam como filtro antes de serem feitas perguntas sobre atitudes, evitando constrangimentos ao tratar assuntos que o respondente desconhece. Em pesquisas que fazem parte de campanhas publicitárias ou de introdução de novas técnicas (p. ex. cuidados da saúde) é necessário verificar conhecimentos além de atitudes e práticas atuais. Para reduzir o nível de ameaça, pode-se iniciar a pergunta com frases como "você sabe por acaso..." ou "a propósito..."

São necessários cuidados para evitar adivinhação por parte de respondentes que não querem admitir falta de conhecimento. Fazendo mais que uma

pergunta reduz-se a possibilidade de acertar a resposta correta por acaso, especialmente em se tratando de assuntos com perguntas que requerem *sim* ou *não*. Usando itens de escolha múltipla, as alternativas precisam ser igualmente plausíveis. É preferível usar perguntas abertas sobre conhecimento de assuntos numéricos (p. ex., Qual a população do Brasil?)

Itens para aferir atitudes e opiniões. Em se elaborando itens para determinar atitudes, há tarefas iniciais. Precisa-se definir claramente o objeto de atitudes, isto é, sobre qual objeto se quer saber algo. Temos: a pessoa X ou as ações da pessoa X ou as filosofias da pessoa X - estas vertentes não são idênticas. É neste ponto que a introdução a um conjunto de itens mostra-se importante. Outro ponto a cuidar é a vertente da atitude que está sendo medida: a afetiva, a cognitiva ou a comportamental. A primeira vertente trata da avaliação de um objeto de atitude: "Considero o atual *playground* um lugar seguro para as crianças da vizinhança". A segunda trata de conhecimentos - certos e errados - acerca de um objeto: "A caixa de areia do *playground* é bem protegida de cachorros?". A terceira trata de ações passadas ou futuras diante de um objeto: "Durante os últimos 15 dias, usei o *playground* X vezes". Como freqüentemente não existe alta correlação entre as três vertentes, deve-se averiguar as três ou aquela que é de interesse principal.

Para verificar a intensidade da atitude, existem dois caminhos fundamentais. Um consiste em fazer uma série de perguntas e, partindo da soma de respostas numa determinada direção, inferir a intensidade da atitude. Em se tratando de atitudes religiosas ou políticas, a alternativa é fazer duas perguntas.

Você se considera mais próximo(a) de que religião:
- Catolicismo 1
- Protestantismo 2
- Que igreja? _____

< seguem outras religiões existentes na população alvo >

Você considera sua fé
- Muito forte 1
- Forte 2
- Mais ou menos 3
- Não tão forte 4
- Fraca 5

No último mês, você freqüentou os cultos
- Quase diariamente 1
- Duas ou três vezes por semana 2
- Uma vez por semana 3
- Semana sim, semana não 4
- Uma vez 5
- Geralmente assiste às festas religiosas 6

A primeira pergunta é mais fatual, embora permita inferências inclusive cognitivas. A segunda é uma pergunta afetiva; a terceira, comportamental. Juntas permitem uma caracterização religiosa do respondente.

Perguntas que contêm dois objetos de atitudes devem ser evitadas. Em vez de perguntar "Você prefere o parque da cidade, que tem campos de futebol, ou os clubes, que geralmente têm bons restaurantes?", devem ser elaboradas no mínimo duas perguntas: "Você prefere o parque da cidade ou os clubes?"; "Entre as seguintes atividades, de qual você gosta mais como lazer num domingo?". A segunda pode ser uma pergunta aberta ou seqüenciada por alternativas como jogar futebol, freqüentar restaurantes, etc. Relaciona-se a este aspecto a questão da pergunta unipolar ou bipolar. Tomando estratégias de enfrentamento como exemplo, poder-se-ia formular dois tipos de perguntas:

duas perguntas unipolares

Geralmente, tento evitar conflito com outras pessoas:
Concordo fortemente 1
Concordo 2
Discordo 3
Discordo fortemente 4
< seguem algumas outras perguntas >
Não levo desaforo para casa:
Concordo fortemente 1
Concordo 2
Discordo 3
Discordo fortemente 4

versus *uma pergunta bipolar*

Tento evitar conflitos com outras pessoas .... 3 2 1 0 1 2 3 ..... Não levo desaforo para casa.

Itens para obter informação fatual. Incluem-se neste grupo perguntas sócio-demográficas: sexo, idade, escolaridade, renda, moradia, etc. Estas perguntas deverão ser vistas no fim da pesquisa, não supondo que tais itens não sejam potencialmente delicados (idade, renda, nível escolar, entre outros). Igualmente, não há como supor que os respondentes, mesmo querendo, forneçam sempre informações corretas. Vale o lembrete de perguntar apenas o que será utilizado. É importante iniciar a última seção com a justificativa "Concluindo, gostaríamos de fazer algumas perguntas que permitam melhor caracterizar o grupo de pessoas com as quais falamos nesta pesquisa". Caso a interação tenha sido demorada ou esta afirmação não tenha sido feita no início da pesquisa, convém reafirmar "Lembramos que todas as suas declarações serão tratadas de maneira confidencial. Os resultados serão apresentados de maneira a não permitir a identificação de participantes individuais". Para ajudar a memória e/ou assegurar respostas acuradas, as perguntas devem ser o mais específicas possível: "Quantos anos fez no seu último aniversário?" vs. "Quantos anos tem?"; "Durante os últimos seis meses, você teve oportunidade de trabalhar com processamento de texto?" vs. "Você tem experiência em computação?"

4.3 - Considerações Técnicas

Voltamos à relação *conceito—item* da Figura 10-1. Partindo dos objetivos, formulam-se perguntas a serem respondidas por meio da pesquisa. As perguntas são transformadas operacionalmente em variáveis e indicadores, apresentadas ao respondente em forma de itens. Desta maneira, é perpassando os itens que se estabelece a relação entre o objetivo de uma pesquisa e os conceitos pesquisados, enquanto as respostas representam o grau de conceituação que o respondente tem acerca do assunto sob investigação.

Aproximamo-nos de uma definição de medição: estabelecer correspondência entre eventos e símbolos, comumente numerais, de tal maneira isomórfica que a variação entre os símbolos corresponda, geralmente de modo linear, à variação entre os eventos. No caso da pesquisa social, *evento* quer dizer "idéias, sentimentos, planos, crenças, bem como origem social, educacional e financeira", aquilo que na definição de Fink e Kosecoff (vide p. acima) está sendo coletado num *survey*, enquanto *símbolo* é a apresentação de alternativas nos itens do instrumento (no caso de perguntas fechadas) ou da codificação das respostas a perguntas abertas. Para que tal correspondência ou medição seja fiel, há que atentar para três aspectos: erro, singularidade e representação. O evento é identificado corretamente e discriminado de outros eventos próximos? Cada evento é representado por apenas um símbolo, e cada símbolo representa apenas um evento? Quais as maneiras com que símbolos (numerais) representam eventos?

Nas ciências sociais, eventos e símbolos se diferenciam em quatro níveis de correspondência, i.é, entre quatro níveis de medição ou escalas sumarizados na Tabela 10-1. A seguir apresentamos exemplos de itens para cada um destes quatro níveis de mensuração.

Tabela 10-1. Escalas de números e suas correspondências

Tipo de Escala	Características da Escala	Exemplos	Características Formais
Nominal	Números ou símbolos são utilizados somente para identificar pessoas, objetos, ou categorias	Placas de carro, Cor de cabelo, Local de nascimento, Estado civil	Equivalência ' = '
Ordinal	Caraterísticas podem ser ordenadas numa dimensão subjacente	Ordem de chegada, Ordem de preferência, Status social, Escala de Likert	além da anterior um item maior do que o outro '>'
Intervalar	Características não somente podem ser ordenadas numa dimensão subjacente, mas intervalos têm tamanho conhecido e podem ser comparados	Escalas de Thurstone, Escala de Likert, Estimativas de distâncias, Temperatura em °C	além das anteriores operações aritméticas nas diferenças entre os números representando eventos
Razão	Além das características da escala anterior, ainda existe um ponto zero absoluto	Salário, Tamanho, Tempo gasto com uma tarefa,	além das anteriores operações aritméticas nos próprios números

Características de Escalas nas Ciências Sociais (Pasquali, 1997; Siegel, 1975; Sommer & Sommer, 1997)

Escala Nominal. Neste tipo de escala utilizam-se números ou símbolos somente para identificar pessoas, objetos ou categorias. Exemplos para as ciências sociais seriam local de nascimento, sexo, estado civil ou atributos como cor de cabelo ou uso de aparelhos como óculos ou bengala. A forma de apresentar estes itens é a seguinte:

Qual o estado civil de V.Sa.?
Solteiro(a) 1
Casado(a) 2
Vivendo maritalmente 3
Desquitado(a) 4
Divorciado(a) 5
Separado(a) 6
Viúvo(a) 7
Outro 8

Apontamos para alguns aspectos deste item. Mesmo ao se preparar um instrumento para auto-aplicação, deve-se pensar em um diálogo com o respondente. Contrariamente a uma declaração de renda ou ficha de procura de emprego, convém estabelecer um bom relacionamento com o respondente. A frase 'Qual o estado civil de V.Sa?' soa melhor do que solicitar simplesmente 'Estado Civil'. Dependendo da população alvo, um maior ou um menor número de alternativas é apropriado: freqüentemente as alternativas 'solteiro, casado, outro' são suficientes. O importante é que as opções (a) sejam mutuamente exclusivas e (b) cubram todas as alternativas. Outra maneira de formular alternativas do estado civil é:

Nunca foi casado(a) 1
Sempre foi casado(a), i.é, casado(a) e nunca divorciado(a) 2
Divorciado(a) 3
Casado(a) novamente 4

Dependendo do objetivo da pesquisa, o primeiro ou o segundo exemplo do item 'estado civil' pode ser mais apropriado. Da mesma forma que a reação inicial do leitor ao segundo exemplo pode ser de estranheza, a maioria dos respondentes a um eventual uso deste item pode assim reagir. É um exemplo concreto de distinção entre o conceito subjacente a ser analisado numa determinada pesquisa (i.é, as quatro categorias do segundo exemplo) e o que pode, do ponto de vista prático e conceitualmente factível ser perguntado à maioria dos respondentes.

Escala Ordinal. Numa escala ordinal, além de identificarem pessoas, objetos ou categorias, números ou símbolos os ordenam numa dimensão subjacente. Exemplos para as ciências sociais seriam hierarquização de preferência ou importância entre pessoas ou objetos, *status* social ou ordem de chegada[5]. A forma de apresentar os itens é:

5 Alternativas nos itens de uma escala Likert são outro exemplo, mas serão tratadas separadamente.

Como você sabe, a Prefeitura está lançando um programa de opções de esporte para os adolescentes deste bairro. Entre as opções que apresento, indique qual deve ser realizada primeiro, qual a segunda, qual a terceira e qual a quarta:

	Número de ordem de importância
Campo de futebol	_____
Área de *skate*	_____
Campo de basquete	_____
Campo de vôlei	_____
Outros, quais?	

A tarefa do respondente é escrever a ordem de importância de realização no espaço indicado. Para cada um dos quatro itens (posteriormente, quatro variáveis) pode-se determinar uma distribuição de freqüência: quantas vezes 'campo de futebol' foi primeira, segunda, terceira e quarta escolhida. A partir disso infere-se sua importância. Igualmente, quais as distribuições para áreas de *skate*, campo de basquete e campo de vôlei? É possível sumarizar os dados indicando quantas vezes cada um dos itens foi mencionado como o mais importante ou qual o valor mediano das menções de importância de cada uma das quatro alternativas. Concluindo: os valores modais e medianos podem ser calculados; a média, não.

Escala Intervalar. Numa escala intervalar, as características não somente podem ser ordenadas conforme uma dimensão subjacente, mas os intervalos entre as alternativas têm tamanho conhecido e podem ser comparados. No caso de julgamentos acerca de eventos pessoais ou sociais (p. ex., satisfação), determinar o tamanho dos intervalos é problemático. Exemplo clássico de uma escala intervalar é a utilizada por Milgram (1974) para determinar o grau de obediência às instruções dos participantes. Ostensivamente, o participante aplicava choques elétricos que variavam entre 15 e 450 volts. O grau de obediência correspondia à voltagem em que o participante se recusava a continuar aplicando mais choques, isto é, quanto mais baixa a voltagem, menos obediente. Mais recentemente, Silva (1999) utilizou a adaptação de um velocímetro (vide Figura 10-2) para o grau de concordância com afirmações numa escala de 0 a 100 por cento.

Figura 10-2. Velocímetro de concordância

Escala de Razão. Exemplos de escalas de razão utilizadas nas ciências sociais são salário ou tempo gasto com uma tarefa. A apresentação dos itens reverte a perguntas abertas:

> Considerando seu tempo livre e de recreação, solicitamos que indique:
> V.Sa. é membro de algum clube esportivo?
> ..... Sim Não
> Caso *sim*,
> – passa quanto tempo por semana nesse clube, em média? ..... _____ horas
> – quanto gasta em atividades no clube, além da mensalidade (em média/mês)? ..... R$ _____

Neste exemplo, a primeira resposta (sim ou não) representa uma medição nominal, enquanto as seguintes representam medições em escala de razão.

Escala Likert. Esta mensuração é mais utilizada nas ciências sociais, especialmente em levantamentos de atitudes, opiniões e avaliações. Nela pede-se ao respondente que avalie um fenômeno numa escala de, geralmente, cinco alternativas: *aplica-se totalmente, aplica-se, nem sim nem não, não se aplica, definitivamente não se aplica.* As afirmações podem ser auto-referentes: "Eu considero importante ter uma área de lazer perto de casa". Ou heterorreferentes: "É importante para uma comunidade ter uma área de lazer". Dependendo do tema subjacente, as alternativas podem, além da dimensão 'aplica-se", seguir dimensões como: 'bom - ruim' ou 'concordo - discordo'. Muitas vezes a dimensão utilizada é apenas uma conseqüência da reformulação do estímulo/item. Avaliam-se objetos ou ações como bons ou ruins. A avaliação de objetos aplica-se ao respondente, ou concorda-se que objetos ou ações têm uma determinada característica: "As oportunidades de lazer na cidade são: boas ... ruins" vs. "Existem oportunidades de lazer nesta cidade" — concordo ... discordo. Convém formular as perguntas de um conjunto de itens de maneira que seu conjunto possa ser respondido na mesma dimensão (veja Sommer, 1991).

> Inicialmente gostaríamos de saber o que os adolescentes deste bairro acham sobre as opções de lazer oferecidas pela Prefeitura. Para cada opção, avalie: *(1) muito ruim, (2) ruim, (3) razoável, (4) bom ou (5) muito bom.*
> Para isto, faça um *círculo* em volta do número que melhor representa sua avaliação:
> Campo de futebol:
> Muito ruim 1
> Ruim 2
> Razoável 3
> Bom 4
> Muito Bom 5
> *<seguem-se os demais itens>*

Conforme mencionado anteriormente, uma série de itens tratando de um mesmo assunto recebe uma introdução comum orientando o respondente: apresenta a temática do conjunto, informa quanto às alternativas e dá instruções concretas, p. ex., 'faça um *círculo* em volta do número que melhor representa sua avaliação'. Geralmente se usam quatro ou cinco alternativas nas escalas tipo Likert, embora se encontrem também itens com duas, três, ou até nove alternativas[6]. Neste exemplo, foram utilizadas cinco alternativas, um número *ímpar*. Uma decisão importante diz respeito ao número par ou ímpar de alternativas, i.é, deixar para o respondente a opção de não se comprometer, podendo marcar um ponto neutro no meio de uma escala com número ímpar de alternativas. Há que diferenciar entre não saber opinar sobre um tema e não querer se comprometer. Quando existe a possibilidade do respondente não ter condições de responder, deve-se deixar a alternativa explícita 'não sei'. Posteriormente, entretanto, tal alternativa não deverá ser tratada como ponto neutro no meio da escala, uma vez que 'indefinido = não saber' é diferente de uma atitude 'indefinido = mais ou menos' no meio de uma escala. Quanto à segunda possibilidade - não querer se comprometer - o respondente provavelmente deixaria o item em branco, não sendo conveniente estimular esse comportamento apresentando a alternativa 'mais ou menos'.

Independentemente do número de alternativas utilizadas, é importante que estejam balanceadas. No exemplo acima, existem duas alternativas positivas e duas negativas. O ponto do meio é 'razoável', isto é, nem positivo nem negativo. Não aceitável seria um grupo de alternativas como 'excelente, muito bom, bom, razoável, ruim' ou 'péssimo, muito ruim, ruim, razoável, bom'. Ambos os exemplos provocam uma avaliação enviesada, em direção positiva ou negativa. Mesmo sem uma alternativa do meio, i.é, numa escala com um número *par* de alternativas, há que assegurar um balanceamento das alternativas.

Considerando o conjunto de itens que compõem um escala tipo Likert, é importante que parte dos itens seja invertida de tal maneira que ora 'concordo' (bom, aplica-se) ora 'discordo' (ruim, não se aplica) represente atitude favorável nos dois itens da escala de ambiente de trabalho (EAT) de Moos (1987):

> A seguir você encontrará uma série de afirmações a respeito do seu *ambiente de trabalho*. Solicito que indique quais afirmações se aplicam a você e quais não se aplicam. Caso se aplique a você, faça um círculo em volta da palavra *SIM*. Caso não se aplique a você, faça um círculo em volta da palavra *NÃO*.
> O trabalho realmente apresenta desafios ..... Sim Não
> Muitas pessoas parecem deixar o tempo passar ..... Sim Não

6 Bortz & Döring (1995, p. 167) fazem referência a um estudo de Matell & Jacoby (1971) em que é argumentado que o número de alternativas não influencia a fidedignidade nem a validade da escala, embora seja necessário lembrar a capacidade discriminatória do respondente.

A concordância 'sim' com o primeiro item e a discordância 'não' com o segundo implicam uma atitude favorável à dimensão 'envolvimento com o trabalho'.

4.4 - Considerações Estatísticas

A diferenciação entre os quatro níveis de escala tem conseqüências importantes quanto à complexidade da análise estatística possível. Dados obtidos em qualquer das escalas podem ser apresentados por meio de estatística descritiva, i.é, tabelas e gráficos. Para utilizar estatísticas inferenciais, que permitem ao pesquisador verificar até que ponto determinadas relações ou diferenças são sistemáticas ou não, há que observar que dados baseados em escalas nominais e ordinais podem ser trabalhados com testes estatísticos não paramétricos. Dados oriundos de escalas intervalares e de razão vão permitir, além de estatísticas não paramétricas, procedimentos paramétricos.

À medida que os testes paramétricos são mais poderosos, permitindo inferências mais complexas, é crucial a questão de a escala Likert poder ser considerada ordinal ou intervalar. Bortz e Döring (1995, p. 168) afirmam que "a controvérsia acerca deste tema tem longa tradição e parece não haver sido resolvida até hoje". Os puristas podem argumentar, com razão, que as alternativas numa escala Likert representam apenas uma medição em nível ordinal. Os valores numéricos (p. ex., 1, 2, 3 e 4) associados às alternativas 'discordo fortemente', 'discordo', 'concordo' e 'concordo fortemente' não permitem operações formais além de '>'. Do ponto de vista prático, pode-se argumentar que a variabilidade nos intervalos não afeta o poder inferencial de uso de estatísticas paramétricas com dados da escala Likert. Importa salientar que esta flexibilidade não se estende à interpretação de médias baseadas em intervalos variáveis. Em outras palavras: para ser cauteloso é apropriado utilizar, para fins descritivos, moda e mediana em lugar da média; e estatísticas paramétricas para fins inferenciais.

Do ponto de vista da análise estatística, medições em nível nominal freqüentemente podem ser convertidas em escalas intervalares. Quando existem apenas duas alternativas, codificadas como '0' e '1', não há necessidade de operações adicionais. Exemplos são perguntas solicitando respostas como sim vs. não, presente vs. ausente, ou sexo. Já itens oferecendo mais de duas alternativas, p. ex. estado civil, região de nascimento, afiliação religiosa, podem ser convertidos em uma série de alternativas binárias através do processo de codificação *dummy*, permitindo operações estatísticas reservadas a escalas intervalares e de razão (veja, p. ex., Tabachnick & Fidell, 1996, ou Stevens, 1986).

5 - Diferenças nos Instrumentos

Até este ponto tratamos do desenvolvimento de um instrumento para *survey* como se fosse independente da maneira de aplicação, i.é, da interação pesquisador - respondente. Após considerações gerais sobre esta interação, trataremos separadamente de entrevistas pessoais, entrevistas por telefone, aplicação de questionários pelo telefone e via internet.

5.1 - Apresentação dos Itens

A apresentação dos itens de um *survey* pode ser conceitualizada como um estímulo de que se espera alguma resposta, algum comportamento, que por sua vez precisa ser de alguma maneira registrado para poder ser analisado. Desta maneira, há potencialmente três atores envolvidos direta ou indiretamente: quem administra o instrumento, quem responde ao instrumento e quem transcreve a informação registrada no instrumento para o processamento e a análise dos dados. Enquanto o objetivo da pesquisa é verificar e analisar variações na resposta, devem ser minimizadas a variabilidade no comportamento de quem responde, a variabilidade atribuível a quem e/ou como se administra o instrumento e a maneira da transcrição das respostas. Vantagens e desvantagens das diferentes formas de aplicação de instrumentos de *survey* relacionam-se diretamente ao poder de minimizar a variabilidade indesejada e ressaltar a variabilidade desejada. A Tabela 10-2 sumariza esses inter-relacionamentos.

		Aplicação do Estímulo: Controle da variabilidade na aplicação do instrumento	
		Baixo	**Alto**
Transcrição da Resposta: Controle da variabilidade na transcrição das respostas ao instrumento	Baixo	Entrevista Pessoal	Questionário enviado via correio ou aplicado em grupo
	Alto	Entrevista via Telefone	Questionário enviado via e-mail / internet.

Tabela 10-2. Formas de aplicação de instrumentos: vantagens e desvantagens

Antes de comentar os quatro modos de apresentar o instrumento de *survey*, seguem-se algumas considerações.

Estimulação concorrente. No caso do instrumento auto-aplicado, é impossível controlar o ambiente onde o respondente preenche o questionário. Já numa interação pessoal pode-se controlar - até certo ponto - a estimulação concorrente pela escolha do local. Não se deve esquecer que o comportamento do entrevistador pode representar uma estimulação concorrente: imagina-se-o manuseando uma prancha com o instrumento, lápis, três fotos - dentre as quais o respondente deve escolher uma - mais o material usado, além daquele a ser usado. Se o entrevistador não for bem treinado, correrá o risco de confundir o respondente antes de obter alguma informação válida. Escolhendo para a aplicação um local calmo, de acesso restrito, com uma boa mesa, reduzem-se interferências indesejadas.

Pessoas envolvidos na administração de survey. Quanto aos atores envolvidos na administração de um instrumento de *survey*, ainda se considera o seguinte: o primeiro ator, que apresenta o instrumento ao respondente no contexto de entrevistas, precisa ser bem treinado para assegurar que a estimulação seja a mais semelhante possível em todos os contatos com os respondentes. A opinião emitida pelo respondente deve representar sua reação às alternativas apresentadas, não a quem as apresentou. Dentro de certos limites, isto pode até ser automatizado quando os itens são apresentados via computador, ou gravados, no caso de entrevistas por telefone. Obviamente, quanto mais estandardizada a apresentação dos estímulos, i.é, dos itens, mais se perde o elemento humano de uma interação, aspecto que leva em conta a situação e o estado de espírito da situação (vide Krosnick, 1999). A preocupação

com uma maior estandardização da apresentação dos itens acontece em levantamentos de dados que: (a) se assemelham a testes, (b) solicitam informações mais objetivas ou (c) coletam dados entre muitos respondentes que precisam ser apurados de maneira rápida.

Considerando o segundo ator (o respondente), a maneira de apresentar os estímulos, itens, deve corresponder às suas habilidades, sejam intelectuais (saber ler) ou físicas (ver, ouvir, discriminar cheiro ou gosto). O que foi dito a respeito da compreensão da linguagem acima estende-se ao uso de símbolos e fotografias. O ☺ é entendido e interpretado como 'concordância'? Aquela foto, caso escolhida pelo respondente como representando um escritório mais confortável, permite a inferência de que o respondente é dinâmico?

Quanto ao modo de registrar as respostas de um *survey* convém pensar, desde o planejamento da pesquisa, no processamento e na análise dos dados. Enquanto respostas a perguntas abertas precisam ser decifradas, transcritas, codificadas, digitadas e verificadas quanto à consistência face às demais respostas (a proverbial mulher de 12 anos que relatou dois abortos e três gravidezes), o uso de um computador na apresentação dos itens e no registro das respostas facilita a apuração e assegura maior fidedignidade aos dados. Questionários que contêm apenas perguntas objetivas podem ser acompanhados de um cartão especial para registro das respostas, que por sua vez pode ser lido mecanicamente. No caso da transcrição por alguém dos dados registrados numa folha de respostas, ou no próprio questionário, deve-se pensar nas capacidades de quem transcreve ou digita. Antes do instrumento ser entregue ao digitador, deve ter sido 'limpo' de tal maneira que não requeira julgamento adicional por parte dele (p. ex., o respondente marcou um 3 ou um 4 naquele item). O "lay-out" do questionário deve permitir orientação no que diz respeito à seqüência da informação a ser transcrita. Se há texto como resposta a perguntas abertas, não somente deve ser legível, mas também claro. Outra questão refere-se ao que deve ser transcrito: o texto todo? apenas uma parte? que parte?. O "lay-out" e as instruções ao respondente devem facilitar a leitura das respostas pelo digitador. No caso de itens de escolha múltipla, devem ser apresentados números em vez de palavras ou letras e pedir que o respondente os *circule* em vez de *marcar com X*.

5.2 - Entrevista Individual

Do ponto de vista da estandardização das perguntas e do potencial para transcrever as respostas, a aplicação pessoal de instrumentos é a mais problemática. Além de exigir treinamento para os aplicadores e para as pessoas que transcrevem as respostas (especialmente a perguntas abertas), a entrevista pessoal é o método mais demorado e mais caro. Sua vantagem é permitir acesso a informações mais delicadas, à parte ser indispensável na fase inicial, estudo-piloto de qualquer tipo de procedimento.

Em instrumentos auto-aplicados pode-se trabalhar com imagens ou apresentar várias alternativas a uma pergunta. No caso de entrevistas, tal uso é mais complicado. Estímulos visuais podem ser preparados para apresentação repetida a respondentes em entrevistas pessoais. Já no caso de entrevistas por telefone, as alternativas precisam ser curtas para que os respondentes não tenham dificuldade de lembrá-las.

5.3 - Questionário Auto-aplicado via Correio ou em Grupo

Do ponto de vista da estandardização das perguntas, questionários auto-aplicados reduzem essa fonte da variabilidade. No que se refere à transcrição das respostas, depende da proporção de perguntas abertas. A desvantagem mais citada de *survey* por correio é a taxa de resposta. Dillman (1972, 1978; Dillman, Christenson, Carpenter & Brooks, 1974; Dillman & Frey, 1974) apresenta uma série de procedimentos que se têm mostrado eficazes para assegurar uma taxa de devolução acima de 50 por cento. Por outro lado, Krosnick (1999) cita pesquisas mais recentes que sugerem que baixas taxas de resposta significam não necessariamente baixo grau de representatividade, especialmente no caso de amostras probabilísticas (vide também Fraser-Robinson, 1991).

5.4 - Entrevista Pessoal via Telefone

Do ponto de vista da estandardização das perguntas e do potencial para transcrever as respostas, a entrevista por telefone - especialmente com apoio de computador - tem grande valor. Embora também precise do treinamento dos entrevistadores, reduz-se consideravelmente o uso de papel, visto que as perguntas são apresentadas na tela do computador para o entrevistador, que as lê para o entrevistado. A seqüência de perguntas pode ser programada de forma que, dependendo da resposta, uma ou outra pergunta seja indicada para ser a próxima. Admitindo que nem toda a população tem acesso a telefone, é preciso atentar para a representatividade da população alvo e da amostra atingida. Em 1988, porém, Rodrigues e colaboradores conseguiram utilizar esta técnica com sucesso no Brasil (vide também Lavrakas, 1993, 1998).

5.5 - Questionário Auto-aplicado via E-mail e Internet

Do ponto de vista da estandardização das perguntas e do potencial para transcrever as respostas, instrumentos distribuídos por meio de e-mail têm grande potencial. Além do mais, são mais rápidos do que *survey* por telefone e mais baratos, porque eliminam custos de entrevistador (*survey* pessoal ou por telefone), papel, impressão, selo (*survey* pelo correio). A problemática de amostragem inerente ao uso do telefone para a coleta de dados é ainda mais séria no uso de e-mail e internet: a população alvo atingível é mais restrita. Schaffer e Dillman (1998) relatam um experimento contrastando diferentes maneiras de contato com respondentes, chegando à conclusão de que técnicas utilizados em *survey* por correio são igualmente válidas para *surveys* por e-mail. Desta maneira, este caminho tem grande potencial para populações que têm acesso a e-mail, seja dentro de uma organização, seja por outras características comuns.

Conclusão

A presente capítulo tratou de bases conceituais de questionários, da estrutura lógica, dos elementos e do contexto social da aplicação do instrumento.

Concluiu-se o texto com considerações acerca de diferentes formas de questionários conforme a maneira de sua aplicação: entrevista individual, pelo telefone, por correio convencional ou eletrônico.

Referências

Aday, L. A. (1989). *Designing and conducting health surveys*. San Francisco: Jossey-Bass.

APA - American Psychological Association. (1992). Ethical principles of psychologists and code of conduct. *American Psychologist, 47*, 1597-1611.

Babbie, E. (1992). *The practice of social research*, 6th ed. Belmont, CA: Wadsworth.

Bingham, W. V. D., & Moore, B. V. (1934). *How to interview*. New York: Harper Collins.

Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation* (2. Aufl.). Berlin: Springer.

Dillman, D. A. (1972). Increasing mail questionnaire response in large samples of the general public. *Public Opinion Quarterly, 36*, 254-257.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. New York: Wiley.

Dillman, D. A., Christenson, J. A., Carpenter, E. H., & Brooks, R. M. (1974). Increasing mail questionnaire response; A four state comparison. *American Sociological Review, 39*, 744-756.

Dillman, d., & Frey, J. H. (1974). Contribution of personalization to mail questionnaire response as n element of a previously tested method. *Journal of Applied Psychology, 59*, 297-301.

Fink, A., & Kosecoff, J. (1985). *How to conduct surveys: A step-by-step guide*. Beverly Hills: Sage.

Fowler, F. J. (1998). Design and evaluation of survey questions. Em L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 343-374). Thousand Oaks, CA: Sage.

Frankel, M. (1983). Sampling theory. Em P. H. Rossi, J. D. Wright, & A. B. Anderson (eds.), *Handbook of survey research* (pp. 21-67). Orlando, FL: Academic Press.

Fraser-Robinson, J. (1991). *Mala direta eficaz* (K. A. Roque, trad; J. A. Nascimento, revisão). São Paulo, SP: Makron Books do Brasil (original publicado em 1989).

Gouveia, V. V., & Günther, H. (1995). Taxa de resposta em levantamento de dados pelo correio: o efeito de quatro variáveis. *Psicologia: Teoria e Pesquisa, 11*, 163-168.

Günther, H., Brito, S. M. O., & Silva, M. S. M. M. (1989). Relação entrevistador - entrevistado: Um exemplo de técnica de entrevista com grupos marginalizados: Meninos na rua. Psicologia: *Reflexão e Crítica, 4* (1/2), 12-23.

Günther, H. & Lopes, Jr., J. (1990). Perguntas abertas vs perguntas fechadas: Uma comparação empírica. *Psicologia: Teoria e Pesquisa, 6*, 203-213.

Henry, G. T. (1990). *Practical sampling*. Thousand Oaks, CA: Sage.

Henry, G. T. (1998). Practical sampling. Em L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 101-126). Thousand Oaks, CA: Sage.

Judd, Ch. M., Smith, E. R., & Kidder, L. H. (1992). *Research methods in social relations*, 6th ed. Fort Worth, TX: Holt, Rinehart & Winston.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Kish, L. (1987). *Statistical design for research*. New York: Wiley.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537-567.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Thousand Oaks: Sage.

Lavrakas, P. J. (1998). Methods for sampling and interviewing in telephone surveys. Em L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 429-472). Thousand Oaks, CA: Sage.

Mangione, Th. W.. (1998). Mail surveys. Em L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 399-427). Thousand Oaks, CA: Sage.

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. *Educational and psychological measurement, 31*, 657-674.

Milgram, S. (1974). *Obedience to authority: An experimental view*. New York: Harper & Row.

Moos, R. H. (1987). *The social climate scales. A user's guide*. Palo Alto, CA: Consulting Psychologists Press, Inc.

Pareek, U., & Rao, T. V. (1980). Cross-cultural surveys and interviewing. Em H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology: Methodology*, vol. 2 (pp. 127-179). Boston: Allyn & Bacon.

Pasquali, L. (1997). *Psicometria: teoria e aplicações*. Brasília, DF: Editora UnB.

Rodeghier, M. (1996). *Surveys with confidence: A practical guide to survey research using SPSS®*. Chicago: SPSS Inc.

Rodrigues, A., Lobel, S. A., Jablonski, B., Monnerat, M., Corga, D., Diamico, K., Pereira, M., & Ferraz, A. (1988). A imagem do político brasileiro. *Psicologia: Teoria e Pesquisa, 4*, 2-11.

Schaffer, D. R., & Dillman, D. A. (1998). Development of a standard e-mail methodology. *Public Opinion Quarterly, 3*, 378-397.

Schuman, H., & Kalton, G. (1985). Survey methods. Em G. Lindzey & E. Aronson (eds.), *Handbook of social psychology*, 3rd ed., Vol 1, (pp. 635-697). New York: Random House.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording , and context*. New York: Academic Press.

Siegel, S. (1975). *Estatística não-paramétrica para as ciências do comportamento* (trad. A. A. de Farias e E. Nick). São Paulo: McGraw-Hill do Brasil. (obra original publicada em 1956).

Silva, A. V. da. (1999). *Comportamentos de motoristas de ônibus: itenerário urbano, estressores ocupacionais e estratégias de enfrentamento*. Brasília: UnB, Dissertação de Mestrado.

Singer, E., van Hoewyk, J., & Maher, M. P. (1998). Does the payment of incentives create expectation effects? *Public Opinion Quarterly, 62*, 152-164.

Solórzano, I. M. (1991). *Padrões de resposta e taxa de participação em levantamentos de campo: Aplicação ao problema do ruído urbano*. Brasília: UnB, Dissertação de Mestrado.

Sommer, R. (1991). Literal vs. metaphorical interpretations of scale terms: A serendipetous natural experiment. *Educational & Psychological Measurement, 51*, 1009-1012.

Sommer, B., & Sommer, R. (1997). *A practical guide to behavioral research: Tools and techniques* (4th ed.). New York: Oxford University Press.

Stevens, J. (1986). *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Sudman, S. (1983). Applied sampling. Em P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), *Handbook of survey research* (pp. 145 - 194). Orlando, FL: Academic Press.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions*. San Francisco, CA: Jossey-Bass.

Tabachnick, B.G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: HarperCollins.

Williams, B. (1978). *A sampler on sampling*. New York: Wiley.

Yaremko, R. K., Harari, H., Harrison, R. C., & Lynn, E. (1986). *Handbook of research and quantitative methods in psychology*. Hillsdale, NJ: Lawrence Erlbaum.

Zdep, S. M., Rhodes, I. N., Schwarz, R. M., & Kilkenny, M. J. (1989). The validity of the randomized response technique. Em E. Singer & S. Presser (Eds.), *Survey research methods: A reader* (pp. 385-390). Chicago, IL: University of Chicago Press. (Originalmente em *Public Opinion Quarterly*, 1979, *43*(4)).

Desenvolvimento das Rotas de Leitura Fonológica e Lexical em Escolares, e de seu Comprometimento em Disléxicos

Fernando Cesar Capovilla, Elizeu Coutinho de Macedo, Marcelo Duduchi e Roberto Amilton Bernardes Sória

Resumo

O presente capítulo é de interesse a pesquisadores em psicologia, fonoaudiologia e lingüística que trabalham na abordagem de processamento de informações, mais especificamente na teoria de duplo processo na leitura de palavras isoladas, proposto por Morton (1969, 1979, 1989). Descreve o *software* CronoFonos 2.0 por nós desenvolvido que implementa um novo algoritmo para análise automática de latência e de duração de locução, bem como da freqüência e duração de segmentos locucionais como função das variáveis psicolinguísticas lexicalidade, freqüência, regularidade e comprimento, além de variáveis de natureza articulatória. Descreve um estudo piloto em que o *software* analisou com sucesso a leitura dos 192 itens da lista de Pinheiro (1994) por 35 universitários.

1 - O sistema multimídia de locução enquanto teste psicométrico e neuropsicológico

O presente sistema de multimídia permite caracterizar o grau de desenvolvimento e/ou comprometimento de rotas de leitura fonológica e lexical. Tal avaliação é feita com base em parâmetros tais como a latência e a duração da locução, a freqüência e a duração de segmentos locucionais, e a freqüência diferencial de erros como função de lexicalidade, regularidade, freqüência, comprimento e composição acústico-articulatória em crianças de pré-escola a quarta série. Assim, o sistema de multimídia pode ser considerado um *teste de habilidade e desempenho* de leitura em voz alta de itens isolados, em que são avaliados padrões de erro e padrões temporais da pronúncia em voz alta. Padrões de erro incluem regularização e escanção (Lemle, 1991), troca de significado e ausência de resposta. Padrões temporais incluem latência de emissão, duração de emissão, número de intervalos segmentares, diferença entre número de segmentos locucionais (em termos de formantes, cf. Russo e Behlau, 1993) e segmentos ortográficos em nível silábico, e razão entre duração de intervalos segmentares e duração de emissão. Os itens que compõem o teste representam uma amostra do universo psicolinguístico diferindo em lexicalidade (não-palavras e palavras), em regularidade das relações grafema-fonema (regulares, regra,

Concluiu-se o texto com considerações acerca de diferentes formas de questionários conforme a maneira de sua aplicação: entrevista individual, pelo telefone, por correio convencional ou eletrônico.

Referências

Aday, L. A. (1989). *Designing and conducting health surveys*. San Francisco: Jossey-Bass.

APA - American Psychological Association. (1992). Ethical principles of psychologists and code of conduct. *American Psychologist, 47*, 1597-1611.

Babbie, E. (1992). *The practice of social research*, 6th ed. Belmont, CA: Wadsworth.

Bingham, W. V. D., & Moore, B. V. (1934). *How to interview*. New York: Harper Collins.

Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation* (2. Aufl.). Berlin: Springer.

Dillman, D. A. (1972). Increasing mail questionnaire response in large samples of the general public. *Public Opinion Quarterly, 36*, 254-257.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method*. New York: Wiley.

Dillman, D. A., Christenson, J. A., Carpenter, E. H., & Brooks, R. M. (1974). Increasing mail questionnaire response; A four state comparison. *American Sociological Review, 39*, 744-756.

Dillman, d., & Frey, J. H. (1974). Contribution of personalization to mail questionnaire response as n element of a previously tested method. *Journal of Applied Psychology, 59*, 297-301.

Fink, A., & Kosecoff, J. (1985). *How to conduct surveys: A step-by-step guide*. Beverly Hills: Sage.

Fowler, F. J. (1998). Design and evaluation of survey questions. Em L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 343-374). Thousand Oaks, CA: Sage.

Frankel, M. (1983). Sampling theory. Em P. H. Rossi, J. D. Wright, & A. B. Anderson (eds.), *Handbook of survey research* (pp. 21-67). Orlando, FL: Academic Press.

Fraser-Robinson, J. (1991). *Mala direta eficaz* (K. A. Roque, trad; J. A. Nascimento, revisão). São Paulo, SP: Makron Books do Brasil (original publicado em 1989).

Gouveia, V. V., & Günther, H. (1995). Taxa de resposta em levantamento de dados pelo correio: o efeito de quatro variáveis. *Psicologia: Teoria e Pesquisa, 11*, 163-168.

Günther, H., Brito, S. M. O., & Silva, M. S. M. M. (1989). Relação entrevistador - entrevistado: Um exemplo de técnica de entrevista com grupos marginalizados: Meninos na rua. Psicologia: *Reflexão e Crítica, 4* (1/2), 12-23.

Günther, H. & Lopes, Jr., J. (1990). Perguntas abertas vs perguntas fechadas: Uma comparação empírica. *Psicologia: Teoria e Pesquisa, 6*, 203-213.

Henry, G. T. (1990). *Practical sampling*. Thousand Oaks, CA: Sage.

Henry, G. T. (1998). Practical sampling. Em L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 101-126). Thousand Oaks, CA: Sage.

Judd, Ch. M., Smith, E. R., & Kidder, L. H. (1992). *Research methods in social relations*, 6th ed. Fort Worth, TX: Holt, Rinehart & Winston.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Kish, L. (1987). *Statistical design for research*. New York: Wiley.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537-567.

Lavrakas, P. J. (1993). *Telephone survey methods: Sampling, selection, and supervision* (2nd ed.). Thousand Oaks: Sage.

Lavrakas, P. J. (1998). Methods for sampling and interviewing in telephone surveys. Em L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 429-472). Thousand Oaks, CA: Sage.

Mangione, Th. W.. (1998). Mail surveys. Em L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 399-427). Thousand Oaks, CA: Sage.

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study I: Reliability and validity. *Educational and psychological measurement, 31*, 657-674.

Milgram, S. (1974). *Obedience to authority: An experimental view*. New York: Harper & Row.

Moos, R. H. (1987). *The social climate scales. A user's guide*. Palo Alto, CA: Consulting Psychologists Press, Inc.

Pareek, U., & Rao, T. V. (1980). Cross-cultural surveys and interviewing. Em H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology: Methodology*, vol. 2 (pp. 127-179). Boston: Allyn & Bacon.

Pasquali, L. (1997). *Psicometria: teoria e aplicações*. Brasília, DF: Editora UnB.

Rodeghier, M. (1996). *Surveys with confidence: A practical guide to survey research using SPSS®*. Chicago: SPSS Inc.

Rodrigues, A., Lobel, S. A., Jablonski, B., Monnerat, M., Corga, D., Diamico, K., Pereira, M., & Ferraz, A. (1988). A imagem do político brasileiro. *Psicologia: Teoria e Pesquisa, 4*, 2-11.

Schaffer, D. R., & Dillman, D. A. (1998). Development of a standard e-mail methodology. *Public Opinion Quarterly, 3*, 378-397.

Schuman, H., & Kalton, G. (1985). Survey methods. Em G. Lindzey & E. Aronson (eds.), *Handbook of social psychology*, 3rd ed., Vol 1, (pp. 635-697). New York: Random House.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording , and context*. New York: Academic Press.

Siegel, S. (1975). *Estatística não-paramétrica para as ciências do comportamento* (trad. A. A. de Farias e E. Nick). São Paulo: McGraw-Hill do Brasil. (obra original publicada em 1956).

Silva, A. V. da. (1999). *Comportamentos de motoristas de ônibus: itenerário urbano, estressores ocupacionais e estratégias de enfrentamento*. Brasília: UnB, Dissertação de Mestrado.

Singer, E., van Hoewyk, J., & Maher, M. P. (1998). Does the payment of incentives create expectation effects? *Public Opinion Quarterly, 62*, 152-164.

Solórzano, I. M. (1991). *Padrões de resposta e taxa de participação em levantamentos de campo: Aplicação ao problema do ruído urbano*. Brasília: UnB, Dissertação de Mestrado.

Sommer, R. (1991). Literal vs. metaphorical interpretations of scale terms: A serendipetous natural experiment. *Educational & Psychological Measurement, 51*, 1009-1012.

Sommer, B., & Sommer, R. (1997). *A practical guide to behavioral research: Tools and techniques* (4th ed.). New York: Oxford University Press.

Stevens, J. (1986). *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Sudman, S. (1983). Applied sampling. Em P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), *Handbook of survey research* (pp. 145 - 194). Orlando, FL: Academic Press.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions*. San Francisco, CA: Jossey-Bass.

Tabachnick, B.G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: HarperCollins.

Williams, B. (1978). *A sampler on sampling*. New York: Wiley.

Yaremko, R. K., Harari, H., Harrison, R. C., & Lynn, E. (1986). *Handbook of research and quantitative methods in psychology*. Hillsdale, NJ: Lawrence Erlbaum.

Zdep, S. M., Rhodes, I. N., Schwarz, R. M., & Kilkenny, M. J. (1989). The validity of the randomized response technique. Em E. Singer & S. Presser (Eds.), *Survey research methods: A reader* (pp. 385-390). Chicago, IL: University of Chicago Press. (Originalmente em *Public Opinion Quarterly*, 1979, *43*(4)).

Desenvolvimento das Rotas de Leitura Fonológica e Lexical em Escolares, e de seu Comprometimento em Disléxicos

Fernando Cesar Capovilla, Elizeu Coutinho de Macedo, Marcelo Duduchi e Roberto Amilton Bernardes Sória

Resumo

O presente capítulo é de interesse a pesquisadores em psicologia, fonoaudiologia e lingüística que trabalham na abordagem de processamento de informações, mais especificamente na teoria de duplo processo na leitura de palavras isoladas, proposto por Morton (1969, 1979, 1989). Descreve o *software* CronoFonos 2.0 por nós desenvolvido que implementa um novo algoritmo para análise automática de latência e de duração de locução, bem como da freqüência e duração de segmentos locucionais como função das variáveis psicolinguísticas lexicalidade, freqüência, regularidade e comprimento, além de variáveis de natureza articulatória. Descreve um estudo piloto em que o *software* analisou com sucesso a leitura dos 192 itens da lista de Pinheiro (1994) por 35 universitários.

1 - O sistema multimídia de locução enquanto teste psicométrico e neuropsicológico

O presente sistema de multimídia permite caracterizar o grau de desenvolvimento e/ou comprometimento de rotas de leitura fonológica e lexical. Tal avaliação é feita com base em parâmetros tais como a latência e a duração da locução, a freqüência e a duração de segmentos locucionais, e a freqüência diferencial de erros como função de lexicalidade, regularidade, freqüência, comprimento e composição acústico-articulatória em crianças de pré-escola a quarta série. Assim, o sistema de multimídia pode ser considerado um *teste de habilidade e desempenho* de leitura em voz alta de itens isolados, em que são avaliados padrões de erro e padrões temporais da pronúncia em voz alta. Padrões de erro incluem regularização e escanção (Lemle, 1991), troca de significado e ausência de resposta. Padrões temporais incluem latência de emissão, duração de emissão, número de intervalos segmentares, diferença entre número de segmentos locucionais (em termos de formantes, cf. Russo e Behlau, 1993) e segmentos ortográficos em nível silábico, e razão entre duração de intervalos segmentares e duração de emissão. Os itens que compõem o teste representam uma amostra do universo psicolinguístico diferindo em lexicalidade (não-palavras e palavras), em regularidade das relações grafema-fonema (regulares, regra,

e irregulares), em freqüência de ocorrência na língua (alta e baixa), em comprimento (mono, bi, tri, e tetra-silábicas) e em padrão de letra ou fonte (caracteres de forma e cursivos).

Trata-se de um teste construído por referência a um construto teórico. Os *traços latentes* avaliados pelo teste consistem nos construtos *rotas de leitura*. A construção do teste foi feita a partir da *abordagem teórica* de processamento de informações em psicologia cognitiva (Eysenck & Keane, 1990), mais especificamente da *teoria de duplo processo na leitura* (Ellis & Young, 1988; Morton, 1969, 1979, 1989). Tal teoria propõe a existência de duas rotas principais para a leitura: a fonológica e a lexical. A *seleção de itens para a composição do teste* foi feita de modo a *cobrir as principais variáveis psicolinguísticas* (lexicalidade, regularidade, freqüência de ocorrência, comprimento, tipo de letra) em seus níveis representativos (palavra vs. não-palavra, regular vs. regra vs. irregular, alta vs. baixa, bi vs. trisilábicas) *que permitem distinguir as duas rotas*. Para tanto, foi empregada uma lista já publicada de itens psicolinguísticos (Pinheiro, 1994). No entanto, *formas paralelas* são facilmente elaboráveis, já que uma série de outras listas encontram-se disponíveis (Françozo, n.p.).

Assim, embora o presente teste possa ser considerado como um teste de habilidade e desempenho de leitura, seus propósitos ultrapassam a esfera prática, objetivando fornecer algo além de um mero escore geral que permita comparar o desempenho de um respondente em relação ao seu grupo de referência ou de um mesmo respondeste em relação a si próprio em tempos diferentes. Como se trata de um *teste referente a construto*, tem uma sólida *base teórica subjacente* que permite tanto aplicações práticas de diagnóstico quanto a promoção de progresso cumulativo da ciência psicológica. O progresso científico pode ser obtido por meio do uso do teste em pesquisa neuropsicolinguística para identificar os padrões temporais de emissão vocálica típicos dos diferentes quadros disléxicos (Seidenberg & McClelland, 1989), bem como dos vários estágios de desenvolvimento da leitura durante a alfabetização escolar (Frith, 1985), e durante sua educação e reeducação na clínica psicopedagógica (Metz-Lutz, 1993; Serón, 1993). Nesses campos, dadas as dificuldades tecnológicas de registro, a escrita tem sido mais bem documentada do que a leitura. Com o advento do presente teste é possível documentar precisamente os padrões de leitura em voz alta frente a uma ampla amostra psicolinguística de modo a testar e aprofundar o conhecimento teórico já disponível no campo.

Além dos propósitos de pesquisa em neuropsicolinguística, o teste também pode ser empregado para propósitos práticos de diagnóstico clínico, permitindo distinguir entre diferentes tipos de dislexia. Por exemplo, na dislexia periférica do tipo *leitura letra-a-letra* (Pinheiro, 1994; Shallice, 1990) os pacientes tendem a pronunciar a palavra apenas depois de terem construído a sua pronúncia letra-a-letra, ou seja, depois de terem recuperado laboriosamente o som de cada letra e os integrado numa só locução, o que normalmente ocorre de modo subvocal ou sub-audível. Neste caso, em termos de padrões temporais, é esperada uma latência de resposta significantemente maior que a normal, e em clara proporção direta ao comprimento da palavra. Tais pacientes passam a atacar as palavras uma letra por vez porque sua dificuldade está em captar a configuração geral da palavra escrita como um todo. Mas esta estratégia compensatória de ataque letra-a-letra só é possível porque nas letras

de forma há um espaço delimitador entre uma letra e outra. Quando tais espaços são removidos, como por exemplo quando as mesmas palavras são impressas em letras cursivas, o ataque da palavra por leitura letra-a-letra já não é mais possível, e portanto, os pacientes mostram-se incapazes de ler o mesmo material impresso em letras cursivas que já haviam lido em letra de forma. No presente teste, são esperados erros de ausência de resposta e de troca de resposta frente às mesmas palavras escritas em letras cursivas.

2 - Substrato teórico

CRONOFONOS 2.0 (Capovilla, Gonçalves, Macedo e cols., 1995; Duduchi e cols., 1995; Macedo, Duduchi, Soria & Capovilla, n.p.) é um *software* para análise de *latência* e de *duração de locução* como função das variáveis psicolinguísticas lexicalidade, freqüência, regularidade e comprimento. É ainda capaz de indentificar *freqüência* e *duração* de *segmentos locucionais*. Isto é importante para o desenvolvimento de pesquisa orientada pela teoria de duplo processo de leitura de Morton (1979), bem como para o desenvolvimento de instrumentos diagnósticos relacionados à teoria que possam ajudar a compreender quadros de dislexias do desenvolvimento (Seymour & Pinheiro, 1995) e de alexias adquiridas (Lecours, Delgado & Pimenta, 1993). Assim, o *software* pode ser empregado para comparar o número de segmentos locucionais com o de segmentos silábicos ortográficos. Isto permite distinguir entre duas diferentes rotas que podem ser usadas para a leitura em voz alta de palavras isoladas, a fonológica e a lexical. Na rota fonológica a leitura ocorre por decodificação de acordo com regras de correspondência grafema-fonema, e a pronúncia é construída segmento a segmento. Na rota lexical a leitura ocorre por acesso direto ao léxico, sem a intermediação de decodificação fonológica, e a pronúncia é emitida como um todo após recuperação lexical.

A rota lexical pode ser empregada com maior sucesso com palavras de alta freqüência na língua, ou quando houve suficiente exposição visual à palavra escrita, de modo que ambos a pronúncia e o acesso ao significado podem ser obtidos diretamente sem a mediação de processos de decodificação fonológica. Assim, pseudo-palavras, ou seja, seqüências de caracteres às quais não corresponde qualquer significado, que têm freqüência de ocorrência zero, não podem ser lidas lexicalmente, a menos que compartilhem tantos elementos em comum com palavras reais de alta freqüência que possam ser lidas por analogia a estas. Assim, a rota lexical é muito sensível às variáveis freqüência e lexicalidade: quanto maior a freqüência de ocorrência de uma palavra na língua para um dado respondente, tanto maior é a tendência ao uso da rota lexical (ou visual direta) para a leitura desta palavra. Assim, quanto maior a exposição de um dado respondente a uma palavra, tanto maior a representação lexical desta, e tanto menor seu limiar de ativação. Pseudo-palavras que têm freqüência de ocorrência zero não podem ser lidas lexicalmente, a menos que sejam tão parecidas com palavras reais de alta freqüência que possam ser lidas por analogia visual a elas.

Como a leitura pela rota fonológica é baseada no processo de decodificação fonológica de acordo com regras de correspondência grafema-fonema, quanto maior a regularidade nas correspondências grafema-fonema que compõem uma dada palavra, tanto maior a facilidade com que a rota fonológica pode ser empregada para

a sua leitura. Conforme Lemle (1991), há três tipos de relação grafema-fonema: regulares, regra, e irregulares. Relações do tipo *regulares* são aquelas que envolvem apenas correspondência grafema-fonema do tipo biunívoca, ou seja, para cada grafema um fonema, e para cada fonema um grafema. Relações do tipo *regra* são aquelas em que a correspondência grafema-fonema não é biunívoca, mas depende de regras de posição, ou seja, dependendo da posição em que ocupa na palavra, um mesmo grafema pode ser pronunciado diferentemente, e um mesmo fonema pode dever ser escrito com grafemas diferentes. Relações grafema-fonema *irregulares* são aquelas em que não há regras de correspondência biunívoca ou regras de posição, sendo que um mesmo grafema pode ser pronunciado, e um mesmo fonema pode ser escrito, de maneiras diferentes, independentemente de quaisquer regras, sendo que a única maneira de decidir sobre a pronúncia e/ou grafia corretas é por acesso ao dicionário, ou léxico da língua. Assim, a rota fonológica pode ser empregada com facilidade para a obtenção da pronúncia correta de itens (palavras ou pseudo-palavras) regulares, e com uma menor facilidade para a obtenção da pronúncia correta de itens regra. A obtenção da pronúncia correta de palavras irregulares requer o uso da rota lexical, ou seja, o conhecimento por exposição prévia (léxico mental) sobre a maneira correta como elas devem ser pronunciadas e escritas. Em suma, a facilidade com que a rota fonológica pode ser usada para a leitura depende muito do grau de regularidade (i.e., de correspondência nas relações grafema-fonema), mas não depende em nada da lexicalidade. Como a variável lexicalidade não deve afetar a leitura pela rota fonológica, é irrelevante se a seqüência pronunciável de caracteres a ser lida tem ou não significado, ou seja, é irrelevante se ela constitui uma palavra real ou apenas uma pseudo-palavra.

Como a pronúncia pela rota lexical tende a ser emitida após endereçamento e recuperação lexical como um todo não-segmentado, enquanto que a pronúncia pela rota fonológica tende a ser construída segmento a segmento por decodificação fonológica, considerando-se um mesmo comprimento constante (medido tanto em termos de número de grafemas como em termos de número de sílabas componentes), a pronúncia de palavras lidas fonologicamente deve ter maior duração locucional do que a de palavras lidas lexicalmente. Além disso, as locuções de palavras lidas fonologicamente também devem apresentar um maior número de segmentos discerníveis do que as locuções de palavras lidas lexicalmente. Portanto, é possível que a correspondência entre o número de segmentos locucionais discerníveis no espectrograma e o número de segmentos silábicos presentes na ortografia seja maior para itens lidos fonologicamente do que para aqueles lidos lexicalmente. Subtraindo-se o número de segmentos silábicos presentes na ortografia do número de segmentos discerníveis no espectrografia, chega-se a uma medida de correspondência orto-espectrográfica. Um valor zero indicaria correspondência plena, isto é, o número de segmentos discerníveis na espectrografia corresponderia àquele da ortografia; um valor positivo indicaria um maior número de segmentos discerníveis na espectrografia do que na ortografia; e um valor negativo indicaria um menor número de segmentos discerníveis na espectrografia do que na ortografia. *Falsos positivos* poderiam ser obtidos em casos de gagueira; correspondência plena poderia ser obtida em pronúncia na leitura por decodificação estrita como em certos períodos da alfabetização, e *falsos negativos* deveriam caracterizar a pronúncia na leitura em adultos normais competentes, que é usualmente lexical.

É imprescindível que a identificação do número de segmentos discerníveis não seja feita subjetivamente por meio de julgamento de especialistas humanos diante de registros espectrográficos, mas que seja feito precisa, objetiva, imparcial e automaticamente por meio de algoritmos especiais implementados em programas de computador e calibrados para o *corpus* contendo todas as variáveis de interesse igualmente representadas. Isto impede a contaminação não apenas por imprecisões de registro dos juízes como também pelo efeito das expectativas teóricas dos investigadores (Capovilla, 1989; Rosenthal e Rosnow, 1984). O *software* CRONOFONOS 2.0 (Capovilla, Gonçalves, Macedo e cols., 1995; Duduchi e cols., 1995; Macedo, Duduchi, Soria & Capovilla, n.p.) por nós desenvolvido, implementa o sofisticado algoritmo de Lee e Hahn (1994), explicado no apêndice, e corrigido por Soria (n.p.). Seu objetivo é permitir o registro preciso e automático em tempo real do início e do término não apenas de locuções inteiras como também de seus segmentos, com vistas a oferecer uma contagem precisa de segmentos perfeitamente customizável a diferentes *corpora*. Isto é obtido por meio da manipulação de até cinco parâmetros independentes que, uma vez estabelecidos, operam em todo o *corpus* fornecendo uma medida documentada, precisa e objetiva do efeito de cada uma das categorias psicolinguísticas de interesse do examinador para aquele *corpus* específico.

É preciso lembrar, no entanto, que o uso de marcações manuais em pesquisa é mais comum do que se gostaria de admitir, mesmo em pesquisas de ótima qualidade empregando espectrogramas. Por exemplo, um dos estudos acústicos representativos do campo que empregam *software* específico para análise da dinâmica temporal de repetições em crianças pré-escolares com gagueira é o de Throneburg e Yairi (1994). Este estudo empregava o *software CSpeech* (Milenkovic, 1987) para medir, nos *displays* visuais dos espectrogramas, as durações das unidades de repetição faladas, os intervalos de silêncio entre as unidades, e a disfluência total. Aqueles autores descobriram que a duração total das disfluências das crianças que gaguejavam era significantemente mais curta devido aos seus intervalos silenciosos serem mais curtos quando comparados às disfluências de unidades de repetição iguais produzidas pelos sujeitos do grupo controle. Conforme Yairi e Lewis (1984), uma unidade de repetição era definida como a produção de um segmento extra (como, por exemplo, *b-but* e *and-and*) e uma unidade dupla de repetição era definida como duas produções extras do segmento (como *b-b-but*, e *and-and-and*). Para a operação desse *software*, os autores precisavam selecionar as unidades de repetição marcando-as, eles próprios, por meio da colocação de cursores no início e final da curva de energia espectral. CRONOFONOS 2.0 poderia ser empregado para automatizar por completo tal processo, permitindo prescindir de seleção e marcação manual. A propósito, tal preocupação para com a computadorização tem caracterizado a maior parte de nossos trabalhos (Capovilla, 1992, 1993, 1994 a,b; Capovilla, no prelo a,b,c; Capovilla, Capovilla e cols. no prelo a,b; Capovilla, César e cols. 1993; Capovilla, Colorni e cols. 1995; Capovilla, Gonçalves e cols. 1996, Capovilla, Haydu e cols. 1995; Capovilla, Macedo, e cols. 1993 a,b, 1994 a,b, 1995 a,b,c,d, 1996 a,b,c,d,e,f; Capovilla, Nunes e cols. no prelo; Capovilla, Raphael e cols. 1996; Feitosa e cols. 1994; Gonçalves, Capovilla e cols. 1995; Gonçalves, Macedo e cols. 1995 a,b; Macedo e cols. 1994; Thiers e cols. 1994).

As expectativas gerais quanto aos resultados da aplicação de CronoFonos 2.0 que derivam a partir do modelo descrito são claras: Como a pronúncia por leitura lexical de palavras tipicamente irregulares e de alta freqüência tende a ser emitida como um todo não-segmentado após recuperação lexical, é possível que nesta comparação entre segmentos locucionais e ortográficos surja um maior número de *falsos negativos* (i.e., o número de segmentos locucionais detectado é inferior ao número de segmentos silábicos presentes na ortografia). Já, como a pronúncia de palavras lidas por decodificação fonológica tende a ser construída segmento por segmento, é possível que seja identificada uma maior concordância entre a freqüência de segmentos locucionais e ortográficos. Como a leitura fonológica tende a ocorrer em pseudo-palavras regulares, e em palavras longas regulares e de baixa freqüência, então espera-se obter uma maior concordância entre segmentos locucionais e ortográficos nestas. Mais sistematicamente, as expectativas poderiam ser assim sumariadas:

Como as pseudo-palavras tendem a ser lidas fonologicamente sendo a pronúncia construída segmento a segmento, e como em leitores adultos competentes as palavras reais tendem a ser lidas mais lexicalmente, espera-se que a freqüência média de falsos negativos seja inferior em pseudo-palavras do que em palavras reais, e que a duração locucional seja maior em pseudo-palavras do que em palavras reais. Como as palavras e pseudo-palavras regulares podem ser lidas pela rota fonológica com maior precisão do que as regra, e estas do que as irregulares, então espera-se que a freqüência média de falsos negativos nas regulares seja menor do que aquela nas regra, e nas regra menor do que nas irregulares. Como, em termos gerais mas não necessariamente, quanto mais longa a palavra tanto maior o número de segmentos silábicos que a compõem, então espera-se uma maior freqüência média de falsos negativos nos itens longos do que nos curtos, espera-se que a latência seja maior nos itens longos do que nos curtos, e espera-se que a duração locucional seja proporcional ao comprimento dos itens. Como palavras de alta freqüência tendem a ser lidas mais lexicalmente do que aquelas de baixa freqüência, e como a leitura pela rota lexical é mais "monolítica" (i.e., menos segmentada), então espera-se encontrar uma maior freqüência média de falsos negativos em palavras de alta do que naquelas de baixa freqüência, e espera-se que a duração locucional das palavras de alta freqüência seja inferior àquela de baixa freqüência. No entanto, se os respondentes forem adultos competentes, e a diferença de freqüência relativa entre os dois conjuntos de palavras não for grande o suficiente a ponto de ser substancial para esses respondentes, é possível não encontrar efeito de freqüência, em falsos negativos e/ou em duração locucional .

O presente estudo foi elaborado para testar tais expectativas acerca do funcionamento do *software* CronoFonos 2.0. Como o presente estudo confirma as expectativas básicas, ele propõe o emprego futuro do *software* CronoFonos 2.0 para proceder a uma análise computadorizada do uso das estratégias de leitura fonológica e lexical por parte de crianças de pré-escola até a quarta série do primeiro grau. O *software* será programado para apresentar a lista de palavras de Pinheiro (1994) após o balanceamento da mesma e a correção de alguns possíveis desequilíbrios na composição dos conjuntos de palavras de alta e baixa freqüência, conforme identificados no estudo-piloto preliminar aqui relatado. As crianças serão expostas à lista de itens

com diferentes valores das variáveis independentes: lexicalidade, regularidade, freqüência, comprimento, e composição acústico-articulatória. O *software* fará análise dos parâmetros das variáveis dependentes como latência e duração da locução, freqüência e duração de segmentos locucionais, e freqüência diferencial de erro. Os valores de tais parâmetros sob cada uma das variáveis independentes arroladas serão descritos, e suas alterações ao longo das cinco séries escolares serão descritas. Serão testadas 80 crianças, oito meninos e oito meninas da pré-escola até a quarta série, uma vez a cada semestre durante dois anos, num total de quatro testagens por criança. Assim, serão descritas alterações de desempenho tanto entre grupos quanto intra-grupos, permitindo uma apreciação do efeito do desenvolvimento da habilidade de leitura sobre os valores de todos os parâmetros apontados. Tal estudo proposto está em curso e seus resultados serão divulgados em breve.

3 - *Software* CronoFonos 2.0

A primeira versão do *software* CronoFonos (Duduchi, Macedo, Soria, Capovilla, 1995) empregava o algoritmo de Rabiner e Sambur (1975) para a detecção de pontos iniciais e terminais de locução. Essa versão tinha alguns problemas ocasionados pelo fato de que o algoritmo não analisava os dados em tempo real, mas consumia em média 15s para a detecção do início e fim de cada locução. Já na segunda versão isto foi corrigido (Macedo, Duduchi, Soria, Capovilla, n.p.). Outro problema apresentado por aquele algoritmo de Rabiner e Sambur é que ele não possibilitava a detecção de diferentes segmentos dentro de uma locução. Por outro lado, a nova versão de CronoFonos 2.0 emprega o algoritmo de Lee e Hahn (1994), detalhado na seção abaixo. Ainda assim, tal algoritmo teve que ser corrigido por Soria (n.p.) para fins de sua implementação computacional no presente estudo. Adicionalmente, o algoritmo de Lee e Hahn encontrava-se implementado com cinco parâmetros fixos, não passíveis de manipulação. Já na correção de Soria (n.p.) empregada em CronoFonos 2.0, todos os 5 parâmetros podem ser manipulados sistematicamente de maneira independe um do outro, permitindo assim ajustar os parâmetros em função do *corpus* arrolado pelo experimentador. Isto permite estender o teste do algoritmo com diferentes tipos de locutores, variando em sexo, idade, escolaridade, procedência étnica e regional, etc. Os próprios Lee e Hahn, embora afirmando que seu algoritmo é superior aos demais da literatura, reconheceram que ele ainda não estava exaustivamente testado, necessitando essas mesmas testagens ulteriores com *corpora* diferentes que são permitidas pela correção de Soria, embutida em CronoFonos 2.0.

De modo a permitir ao leitor compreender melhor as características relevantes do programa, as seguintes explicações são pertinentes: CronoFonos 2.0 emprega cinco parâmetros. São eles:

1) número de quadros total (NQT),
2) número de quadros para início da locução (NQI),
3) número de quadros para o final da locução (NQF),
4) nível de energia para início da locução (NEI),
5) nível de energia para final da locução (NEF).

Em CronoFonos 2.0, tanto o nível de energia para início da locução (NEI) quanto aquele para seu término (NEF) podem variar entre 1,1 e 50. O início de

locução, em termos de nível de energia, é dado quando a razão do nível de energia em dado momento em relação ao nível de energia do ruído de fundo (ruído branco) ultrapassa um determinado valor, que pode variar de 1,1 até 50. O final da locução em termos de nível de energia é dado quando aquela razão cai abaixo do valor pré-determinado. O algoritmo de Lee e Hahn calcula em primeiro lugar a energia de fundo a fim de possibilitar sua utilização em diversos ambientes desde escritórios até laboratórios e câmaras anecóicas. Esta característica permite ao algoritmo prescindir da necessidade de ambientes extremamente controlados tais como câmaras anecóicas. Após o cálculo da energia desse ruído de fundo, o algoritmo usa um fator *flutuante* (NEI ou NEF) para a detecção do início ou fim da locução. Assim, como a energia de fundo é sempre levada em consideração pelo algoritmo, programas nele baseados podem ser empregados em quaisquer ambientes, uma vez que eles ajustam-se automaticamente aos diferentes níveis de ruído de fundo. Como o nível de energia nas consoantes plosivas varia abruptamente, enquanto que aquele nas fricativas varia de maneira gradual, o ajuste de NEI e NEF passa a ser crítico, havendo uma maior dificuldade de precisão na detecção do início e término de locuções envolvendo fricativas iniciais e finais do que naquelas envolvendo plosivas. Tais valores devem ser buscados empiricamente de acordo com as questões experimentais e o *corpus* sob exame (i.e., as locuções pelos respondentes dos itens de uma dada lista de palavras), e as questões experimentais consideradas. Com base nos dados do estudo-piloto realizado para a redação do presente projeto, a partir de um *corpus* de 192 itens lidos em voz alta por 35 univeristários de ambos os sexos, foi observado que os parâmetros de nível de energia que produziram maior precisão para início e fim de locução de palavras foram NEI e NEF = 2,5.

Em CronoFonos 2.0, o número de quadros total (NQT) pode variar de 1 a 30. O NQT está relacionado à duração total do intervalo mínimo considerado para busca de alteração de nível de energia. Cada quadro tem duração exata de 10 ms, conforme o algoritmo de Lee e Hahn (1994). Portanto, quando NQT= 1 ou 30, o algoritmo busca alterações no nível de energia de 10 em 10 ms ou de 300 em 300 ms, respectivamente. O número de quadros total (NQT) está relacionado com a precisão com que podem ser detectados inícios e fins de locução, ou ainda com o rigor com que um dado intervalo pode ser caracterizado como silêncio ou como locução.

Em CronoFonos 2.0, o número de quadros para início da locução (NQI) bem como o número de quadros para o final da locução (NQF) variam de 1 a NQT. NQI representa o número de vezes em que o nível de energia deve ultrapassar a razão estabelecida em relação ao número total de quadros (NQT) para fins de determinação do início da locução. NQF representa o número de vezes em que o nível de energia deve estar abaixo da razão estabelecida em relação a NQT. Quanto mais próximo for o valor de NQI em relação ao de NQT, tanto maior será o rigor na determinação da detecção do momento exato do início de uma locução; e quanto menor o valor de NQI, tanto menor será a precisão de detecção do início da locução. Quanto mais próximo for o valor de NQF em relação ao de NQT, tanto maior será o rigor na determinação da detecção do término da locução; e quanto menor for o valor de NQF, tanto menor será esta precisão. Portanto, tais valores devem ser buscados empiricamente de acordo com o *corpus* e as questões experimentais consideradas. No presente estudo com um *corpus* de 192 itens lidos em voz alta por cada um de 35

univeristários de ambos os sexos foi observado que os parâmetros que produziram maior precisão para início e fim de locução de palavras foram NQI= 28, NQF = 28, NQT = 30. Portanto, no presente estudo foi estabelecido que para a determinação do início de locução, o nível de energia devia ultrapassar o valor determinado em pelo menos 28 dos 30 quadros; e do mesmo modo, para a determinação de fim de locução o nível de energia devia estar abaixo do valor determinado em pelo menos 28 dos 30 quadros. A adoção de tais parâmetros permite que o período (i.e., intervalo ou duração) de latência registrado se aproxime com bastante precisão do período de tempo realmente despendido em iniciar a locução (i.e., latência), bem como que a duração da locução registrada se aproxime com bastante precisão do tempo real despendido na locução.

Quanto maior for o número de quadros, tanto menor será a precisão na identificação de segmentos de locução. Mas se o número de quadros for sistematicamente diminuído com o objetivo de aumentar a precisão é possível a ocorrência de erro do tipo falso positivo, ou seja, detectar incorretamente locução onde havia apenas ruído. É possível também detectar incorretamente silêncio onde, na verdade, havia locução (não suficientemente duradoura). Na medida em que a sensibilidade aumenta excessivamente (pelo aumento do número de quadros) o programa passa a detectar um número maior de locuções artificiais, consequentemente aumenta o "fatiamento" do período com o aumento no número de segmentos de locução e de silêncio, o que dificulta o discernimento de locuções relevantes (positivas). Fica assim difícil discernir entre positivos e negativos em meio a tantos falsos positivos e negativos. O nível de energia é representado na ordenada enquanto que o número de quadros é representado na abcissa.

Para a detecção de número de segmentos de locuções o número de quadros total deve ser baixo (aproximadamente cinco quadros), pois esses segmentos têm usualmente curta duração. Mas esta redução pode levar à identificação incorreta de ruídos como sendo segmentos (i.e., falso positivo). Ou seja, a redução do número de quadros faz com que a análise seja feita de 50 em 50 ms, o que é um tempo bastante reduzido quando se considera locuções de palavras, mas não quando se considera segmentos dessas locuções. A redução do número de quadros acarreta perda em precisão da localização exata do início ou término de uma locução, pois o critério pode ser satisfeito por um número de quadros menor (5) mas não maior (10 ou 15). No presente estudo foram derivados outros parâmetros a fim de identificar o número de segmentos numa locução com menor perda possível da precisão. Os parâmetros para isto foram NQT = 8, NQI = 4, NQF = 4, NEI = 5, NEF = 5. Tais valores de parâmetros parecem produzir estimativas razoavelmente precisas, entretanto é necessário verificar o efeito da adoção de outros parâmetros com a presente base de dados. Também é muito provável que tais parâmetros tenham que vir a ser modificados quando o programa for usado para analisar padrões de leitura em voz alta de crianças escolares.

O estabelecimento de parâmetros médios (NEI, NEF, NQT, NQI, NQF) ideais para um dado *corpus* permite que desvios forneçam informações relevantes à compreensão do efeito de uma série de variáveis. Dentre estas, destacam-se as variáveis psicolinguísticas lexicalidade, freqüência, regularidade, e comprimento, cujos efeitos foram examinadas com base na mesma lista por Pinheiro (1994) e Capovilla.

Capovilla e Colorni (no prelo). Também destacam-se as variáveis de natureza fonética e articulatória, tais como examinadas na língua portuguesa por Behlau, Tosi e Pontes (1985), Behlau, Pontes, Tosi e Ganança (1988a), Behlau, Pontes, Tosi, Ganança (1988b), Behlau, Pontes, Tosi, Ganança (1988c), Russo e Behlau (1993). CronoFonos 2.0 pode ser empregado como um instrumento de análise dos efeitos dessas diferentes variáveis, e para o teste de diferentes modelos teóricos testáveis a partir delas.

CronoFonos 2.0 pode também ser empregado clinicamente para avaliar o grau de gagueira e o efeito de tratamentos terapêuticos. Procedimentos diferentes vêm sendo empregados na literatura com este objetivo, como por exemplo o de McClean, Levandowski e Cord (1994). Diferentemente de CronoFonos 2.0 que avalia as propriedades sonoras do *output* vocálico dos respondentes, o procedimento empregado por McClean, Levandowski e Cord consistia em registros eletroglotográficos da vibração da laringe, bem como dos movimentos dos lábios e mandíbulas durante emissões vocálicas repetidas e simples. Assim, diferentemente de CronoFonos 2.0, o procedimento de McClean, Levandowski e Cord pode ser considerado bastante trabalhoso. Conforme mencionado na introdução, estudos acústicos que empregam *software* específico (Milenkovic, 1987) para análise da dinâmica temporal de repetições em gagueira como o de Throneburg e Yairi (1994) ainda fazem marcação manual nos *displays* visuais dos espectrogramas. A grande vantagem de CronoFonos 2.0 é que ele permite automatizar por completo tal processo conferindo-lhe maior objetividade, precisão e rapidez. Por outro lado, representativo também da alta tecnologia envolvendo a análise de padrões temporais articulatórios, é o estudo recente de Byrd (1996) que empregou eletropalatografia para medir a redução e a superposição temporal no tempo de articulação de seqüências de consoantes em função do local de articulação, do modo de articulação, e da estrutura da sílaba. Nesta técnica de eletropalatografia, os sujeitos usavam um palato artificial de acrílico fino contendo 96 eletrodos. O palatômetro varria o palato a uma taxa de 100 Hz com um tempo de varredura de 1.7 ms para coletar todos os 96 valores da amostra, e os eletrodos eram calibrados com *software* especial.

4 - Experimentação

O presente estudo-piloto consistiu num primeiro teste da adequação do programa CronoFonos 2.0 em explorar o efeito de variáveis psicolinguísticas no padrão de leitura em voz alta de 35 universitários de ambos os sexos. Para tanto, ele foi programado para apresentar os 192 itens de uma lista de palavras bastante conhecida (Pinheiro, 1994). CronoFonos 2.0 registrava a freqüência de falsos negativos, a latência e a duração da locução como função das variáveis lexicalidade, regularidade, freqüência e comprimento dos itens.

As expectativas gerais quanto aos resultados da aplicação de CronoFonos 2.0 que derivam a partir do modelo descrito poderiam ser assim sumariadas:

1) Efeito de lexicalidade: Como pseudo-palavras tendem a ser lidas fonologicamente (elas tendem a ser usualmente, mas nem sempre necessariamente são, já que pode haver leitura por analogia....) sendo a pronúncia construída segmento a segmento, e como em leitores adultos competentes as palavras reais tendem a ser

lidas lexicalmente,
   1.1 espera-se que a freqüência média de falsos negativos seja inferior em pseudo-palavras do que em palavras reais.
   1.2 espera-se também que a duração locucional seja maior em pseudo-palavras do que em palavras reais
2) Efeito de regularidade: Como as palavras e pseudo-palavras regulares podem ser lidas pela rota fonológica com maior precisão do que as regra, e estas do que as irregulares, então:
   2.1 espera-se que a freqüência média de falsos negativos nas primeiras seja menor do que aquela nas segundas, e estas menos do que aquelas na terceira.
3) Efeito de comprimento: Como em termos gerais (mas não necessariamente) quanto mais longa a palavra tanto maior o número de segmentos silábicos que a compõem, então:
   3.1 espera-se uma maior freqüência média de falsos negativos nos itens longos do que nos curtos
   3.2 espera-se que a latência seja maior nos itens longos do que nos curtos
   3.3 espera-se que a duração locucional seja proporcional ao comprimento dos itens.
4) Efeito de freqüência: Como palavras de alta freqüência tendem a ser lidas mais lexicalmente do que aquelas de baixa freqüência, e como a leitura pela rota lexical é mais "monolítica" (i.e., menos segmentada), então:
   4.1 espera-se encontrar uma maior freqüência média de falsos negativos em palavras de alta do que naquelas de baixa freqüência.
   4.2 espera-se que a duração locucional das palavras de alta freqüência seja inferior àquela de baixa freqüência.

No entanto, se os respondentes forem adultos competentes, e a diferença de freqüência relativa entre os dois conjuntos de palavras não for grande o suficiente a ponto de ser substancial para esses respondentes, é possível não encontrar efeito de freqüência, em falsos negativos e/ou em duração locucional.

4.1 - Método

4.1.1 - Participantes

Trinta e cinco estudantes de graduação em psicologia participaram voluntariamente do experimento. Eles não estavam informados sobre as hipóteses experimentais, e não tinham conhecimento sobre a teoria subjacente.

4.1.2 - Equipamento e Situação experimental

Foi empregada a segunda versão do *software* CronoFonos (Duduchi, Macedo, Soria, Capovilla, 1995), que apresentava a lista de 192 itens de Pinheiro (1994). O *software* era executado em *notebook* marca Acom, modelo Patriot, com microprocessador *pentium* 100 Mhz com 8 Mb de RAM, HD de 810, matriz ativa, e *kit* multimídia de velocidade quádrupla. O programa era executado em Windows 3.11 com resolução de 640 x 480. As palavras apareciam escritas com letra tipo *Times*

New Roman, tamanho 72. As palavras apareciam escritas em letra preta sobre uma janela branca num fundo azul. Foi também empregado um microfone profissional marca Lesson. Toda a coleta de dados ocorreu numa sala comum do Laboratório de Neuropsicolinguística Experimental do Instituto de Psicologia da Universidade de São Paulo. Os participantes sentavam-se a 30 cm da tela do computador que estava no nível dos olhos, e o microfone ficava a 5 cm da boca. Durante toda a sessão os sujeitos sentavam-se defronte ao *notebook* e ao lado do experimentador.

Procedimento

Os participantes eram expostos individualmente a uma única sessão de cerca de 15 min de duração. As instruções pré-experimentais especificavam que eles deveriam ler em voz alta todas as palavras que aparecessem na tela do computador, independentemente de as palavras serem ou não conhecidas. Antes do início da sessão, os participantes eram instruídos a vocalizar seu nome com voz normal. Isto permitia ao experimentador identificar se o nível de energia da locução estava adequado. Ao término da sessão, o experimentador explicava detalhadamente os objetivos do estudo.

4.2 - Resultados preliminares

No presente estudo havia quatro variáveis independentes: lexicalidade (pseudo-palavra, palavra real), freqüência (baixa, alta), regularidade (regular, regra, irregular), e comprimento (4, 5, 6, 7 letras), e três variáveis dependentes: freqüência de falsos negativos, latência, e duração locucional. Trata-se de um delineamento experimental-estatístico bastante sofisticado, uma vez que usa medidas repetidas com células com ns diferentes, e que a variável *freqüência* encontra-se "aninhada" (*nested*) no nível *palavra real* da variável *lexicalidade*. Como nenhum dos pacotes de análise estatística conseguia analisar os dados em tal delineamento complexo, a única maneira de não omitir qualquer dado ou variável foi combinar as variáveis lexicalidade e a nela "aninhada" freqüência, gerando a variável lexicalidade-frequência com três níveis: pseudo-palavra, palavra de baixa freqüência, palavra de alta freqüência. Assim, o delineamento passou a ter três variáveis independentes e três dependentes. O dados foram analisados via *Systat for Windows*.

ANOVA 3x3x4 de medidas repetidas revelou que a freqüência de falsos negativos foi afetada pela lexicalidade-frequência ($F_{[2,150]} = 14,14$; $p = .000$), pela regularidade ($F_{[2,150]} = 16,77$; $p = .000$), e pelo comprimento ($F_{[3,225]} = 7,53$; $p = .000$), bem como pelas interações duplas entre lexicalidade-frequência e regularidade ($F_{[4,300]} = 3,1$; $p = .01$), lexicalidade-frequência e comprimento ($F_{[6,450]} = 4,52$; $p = .000$), regularidade e comprimento ($F_{[6,450]} = 8,2$; $p = .000$), e finalmente pela interação tripla entre lexicalidade-frequência, regularidade, e comprimento ($F_{[12,900]}$ 3,87; $p = .000$). ANOVA 3x3x4 de medidas repetidas revelou também que a latência locucional foi afetada pela lexicalidade-frequência ($F_{[2,156]} = 47,89$; $p = .000$), pela regularidade ($F_{[2,156]} = 4,91$; $p = .000$), e pelo comprimento ($F_{[3,234]} = 7,22$; $p = .000$), bem como pela interação dupla entre lexicalidade-frequência e comprimento ($F_{[6,468]} = 3,08$; $p = .006$). e pela interação tripla entre lexicalidade-frequência, regularidade, e comprimento ($F_{[12,936]} = 3,22$; $p = .000$). No entanto não houve evidência entre interações duplas entre lexicalidade-frequência e regularidade, ou entre regularidade e comprimento. ANOVA 3x3x4 de medidas repetidas revelou também que a duração locucional foi afetada pela lexicalidade-frequência ($F_{[2,168]} = 4,79$; $p = .009$), bem como pelo comprimento ($F_{[3,252]} = 46,97$; $p = .000$). No entanto, não houve efeito de regularidade ou de quaisquer interações entre as variáveis. Tais dados encontram-se representados nas Figuras 1 a 5.

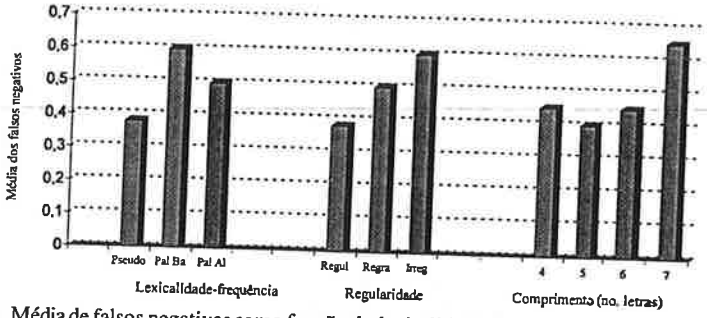


Fig. 11-1. Média de falsos negativos como função de lexicalidade-frequência, regularidade e comprimento do item

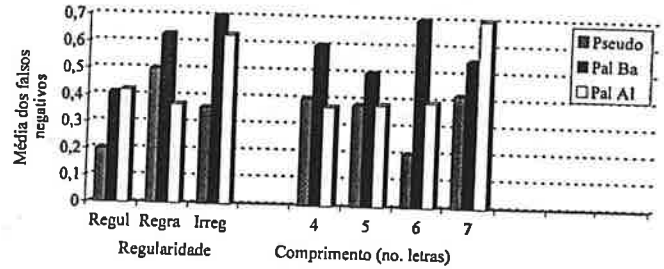


Fig. 11-2. Média de falsos negativos como função de interação dupla entre regularidade e lexicalidade-frequência e da integração dupla entre comprimento (número de letras) e lexicalidade-frequência

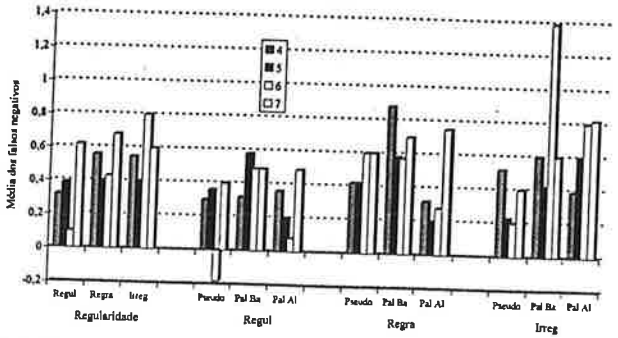


Fig. 11-3. Média de falsos negativos como função de interação dupla entre comprimento (número de letras) e regularidade, e da interação tripla entre comprimento, regularidade e lexicalidade-frequência

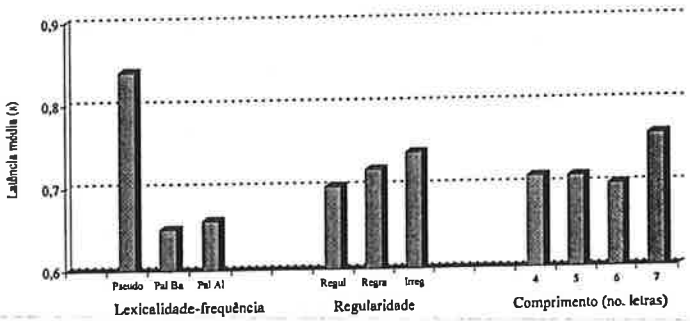


Fig. 11-4. Latência média (s) como função de lexicalidade-frequência, de regularidade e de comprimento do item

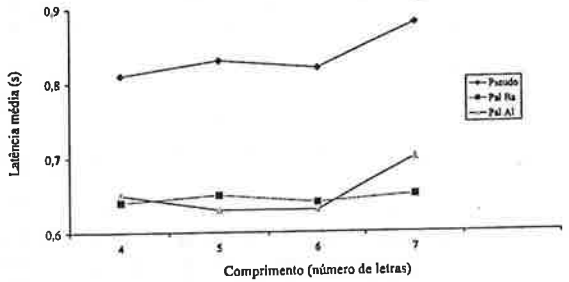


Fig. 11-5a. Latência média (s) como função de interação dupla entre comprimento (número de letras) e lexicalidade-frequência

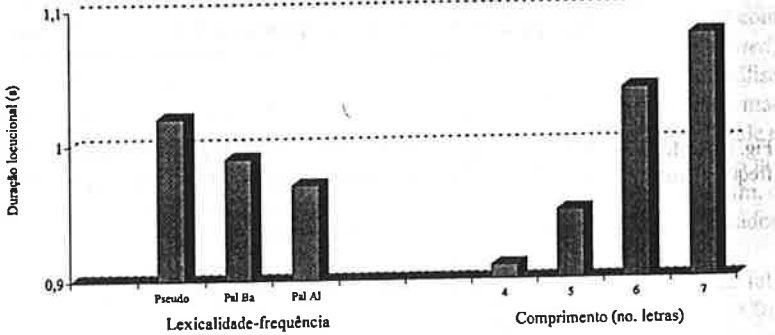


Fig. 11-5b. Duração locucional (s) como função de lexicalidade-freqüência e de comprimento do item (número de letras)

4.2.1 - Freqüência de falsos negativos

Efeitos principais

Em termos de lexicalidade e freqüência, conforme a Figura 1 (à esquerda), como esperado, a média de falsos negativos em pseudo-palavras foi significantemente inferior àquela de palavras tanto de baixa freqüência ($F_{[2,75]}$ = 55,78; p = .000) quanto de alta freqüência ($F_{[2,75]}$= 11,46; p = .000). No entanto, contrariamente às expectativas, a média de falsos negativos em palavras de baixa freqüência foi significantemente superior àquela de alta freqüência ($F_{[2,75]}$= 19,54; p = .000).

Em termos de efeito de regularidade, conforme a Figura 1 (ao centro), como esperado, a média de falsos negativos em palavras regulares foi significantemente inferior àquela de palavras regra ($F_{[2,75]}$ = 14,74; p = .000) e de palavras irregulares ($F_{[2,75]}$ = 44,28; p = .000); e a média de falsos negativos em palavras regra foi significantemente inferior àquela em palavras irregulares ($F_{[2,75]}$ = 4,88; p = .01). Em termos de comprimento, definido como número de letras componentes, o número de falsos negativos foi significantemente superior em itens de 7 letras do que em qualquer um dos demais ($F_{[2,75]}$ = 18,75; 24,13; 19,95; p = .000, .000, .000, para 6, 5 e 4 letras, respectivamente), e em itens de 6 letras do que nos de 5 ($F_{[2,75]}$ = 4,41; p = .016). As demais diferenças não foram significantes.

Efeitos de interação dupla

Em termos de interação entre lexicalidade-frequência e regularidade, conforme a Figura 2 (à esquerda), e conforme esperado, para itens regulares, a freqüência de falsos negativos de pseudo-palavras foi significantemente inferior àquelas de palavras de baixa e de alta freqüência ($F_{[2,75]}$ = 10,97; 17,08; respectivamente, p = .000 para ambas). Não houve diferença entre alta e baixa. O mesmo foi encontrado para itens irregulares ($F_{[2,75]}$ = 43,11 e 11,44 respectivamente, p = .000 para ambas). Para itens regra, a freqüência de falsos negativos em pseudo-palavras foi inferior àquela em palavras de baixa freqüência ($F_{[2,75]}$ = 4,04; p = .022). No entanto, aqui, contrariamente ao esperado, a freqüência de falsos negativos em palavras de baixa foi significantemente superior àquela em palavras de alta ($F_{[2,75]}$ = 19,55; p = .000), e também em pseudo-palavras ela foi maior do que em palavras de alta freqüência ($F_{[2,75]}$ = 8,15; p = .001).

Como esperado, a média de falsos negativos para pseudo-palavras foi menor para as regulares do que para as regra ($F_{[2,75]}$ = 17,65; p = .000) e do que para as irregulares ($F_{[2,75]}$ = 4,21; p = .019); mas maior para as regras do que para as irregulares ($F_{[2,75]}$ = 7,02; p = .002). Também conforme esperado, para palavras de baixa freqüência a freqüência de falsos negativos para regulares foi menor do que para as regra ($F_{[2,75]}$ = 7,9; p = .001) e do que para irregulares ($F_{[2,75]}$ = 16,87; p = .000), e para as regra do que para as irregulares ($F_{[2,75]}$ = 3,85; p = .026). Finalmente, também conforme esperado, para palavras de alta freqüência, a freqüência de falsos negativos para as irregulares foi maior do que aquelas para regulares ($F_{[2,75]}$ = 8,24; p = .001) e regra ($F_{[2,75]}$ = 17,45; p = .000), mas não houve diferença significante entre regulares e regra.

4.2.2 - Latência

Efeitos principais

Em termos de efeito de lexicalidade-frequência, conforme a Figura 4 (à esquerda), e conforme esperado, a latência para pseudo-palavras foi significante-

mente superior àquelas tanto para palavras de baixa quanto de alta freqüência ($F_{[2;78]}$ = 105,52; 110,26; p = .000, .000, respectivamente). No entanto, não houve diferença significante entre as latências de palavras de baixa e alta freqüência.

Em termos de efeito de regularidade, conforme a Figura 4 (ao centro), e conforme esperado, a latência para itens regulares foi significantemente inferior às latências tanto para itens regra ($F_{[2,78]}$ = 5,47; p = .006) quanto para itens irregulares ($F_{[2,78]}$ = 10,85; p = .000). Não houve diferença significante entre as latências para itens regra e aquela para itens irregulares.

Em termos de efeito de comprimento, conforme a Figura 4 (à direita), e de acordo com o esperado, a latência de resposta para itens com sete letras foi significantemente superior àquelas para os demais ($F_{[2,84]}$ = 18,26; 13,6; 6,45; p = .000, .000, .003, para 6, 5 e 4 letras, respectivamente), mas não houve diferenças entre essas.

Interação dupla

Houve interação significante entre lexicalidade-frequência e comprimento na determinação da latência. Conforme a Figura 5 (à esquerda), e como esperado, a latência para pseudo-palavras foi significantemente superior àquelas de palavras reais em todos os níveis de comprimento. A latência para pseudo-palavras foi significantemente maior com 7 letras do que com 6, 5, e 4 ($F_{[2,78]}$ = 9,63; 6,66; 5,17; p = .000, .002. 008, respectivamente). Não houve diferenças entre as latências de pseudo-palavras com 4, 5 e 6 letras. Tampouco houve diferença significante entre as latências de palavras de baixa e alta freqüência para os comprimentos 4, 5 e 6. No entanto, em itens com 7 letras, contrariamente ao esperado, a latência de palavras de alta freqüência foi significantemente superior àquela de baixa freqüência ($F_{[2,78]}$ = 5,94; p = .004).

4.2.3 - Duração locucional

Efeitos principais

Em termos de efeito de lexicalidade-frequência, conforme a Figura 5 (ao centro), a duração locucional de pseudo-palavras foi significantemente superior àquela de palavras de alta freqüência ($F_{[2,84]}$ = 7,78; p = .001), conforme esperado; no entanto não houve diferença significante entre a duração locucional de pseudo-palavras e aquela de palavras de baixa freqüência. Também conforme esperado, foi observada duração locucional significantemente superior de palavras de baixa freqüência do que naquelas de alta freqüência ($F_{[2,84]}$ = 3,87; p = .025).

Em termos de efeito de comprimento, conforme a Figura 5 (à direita), conforme esperado, houve diferenças significantes entre todos os níveis de comprimento, sendo que quanto maior o comprimento, tanto maior a duração locucional. As diferenças de quatro a cinco, de cinco a seis, e de seis a sete foram, respectivamente, $F_{[2,84]}$ = 13,61; 51,05; 10,72, com p = .000 para todos.

4.3 - Discussão e Conclusões

De maneira geral, os dados aqui obtidos adequaram-se bastante bem às expectativas baseadas na literatura. Em primeiro lugar, em termos de lexicalidade, a média de falsos negativos em pseudo-palavras foi significantemente inferior àquela das palavras reais, indicando uma maior segmentação na pronúncia das pseudo-palavras do que na das palavras reais. Além disso, a latência média de pseudo-palavras foi também significantemente superior àquela das palavras reais. Finalmente, a duração locucional de palavras de alta freqüência foi significantemente menor do que a de pseudo-palavras e do que a de palavras de baixa freqüência. Assim, pseudo-palavras produziram significantemente maior latência, maior duração e maior freqüência de segmentação do que palavras reais, o que é plenamente compatível com a interpretação de que tais pseudo-palavras são lidas fonologicamente. Também, a duração locucional de palavras de alta freqüência foi menor do que a das de baixa e do que a das de pseudo-palavras, o que é compatível com a interpretação de que elas são lidas lexicalmente.

Em segundo lugar em termos de regularidade, a média de falsos negativos foi inferior em itens regulares do que nos regra, e nesses do que nos irregulares, indicando uma maior segmentação na pronúncia de itens regulares do que nas de itens regra, e destes do que em itens irregulares. Além disso, os itens regulares (em que não havia discrepância entre as pronúncias por decodificação e aquelas por acesso lexical) produziram latência menor do que os regra e os irregulares. Em terceiro lugar, em termos de comprimento, a média de falsos negativos em itens com 7 letras foi maior do que a dos demais, e a dos de 6 foi maior do que a dos de 5. Consistente com isto, a latência média em itens com 7 letras foi significantemente maior do que a dos demais. Finalmente, em plena consistência, a duração da locução de itens com 7 itens foi significantemente maior do que aquela com 6, de 6 do que de 5, e de 5 do que de 4.

Restam por explicar algumas discrepâncias dos dados presentes em relação aos esperados de acordo com a literatura, especialmente entre a média de falsos negativos em palavras de alta e baixa freqüência. As expectativas baseadas na literatura seriam de que deveria haver maior freqüência de falsos negativos em palavras de alta freqüência do que naquelas de baixa, mas os resultados foram o oposto. Também em contradição com as expectativas baseadas na literatura é a ausência de diferença na latência entre palavras de baixa e alta freqüência. A análise da interação entre lexicalidade-frequência e regularidade revela que os dados de freqüência de falsos negativos estão conforme as expectativas teóricas tanto para palavras regulares quanto para palavras irregulares, nas quais as pseudo-palavras produziram menos falsos negativos do que as palavras reais, tanto de baixa quanto de alta freqüência. Para as palavras regra, a freqüência de falsos negativos também foi menor para pseudo-palavras do que para palavras regra. Todo o problema concentrou-se nas palavras regra de alta freqüência, em que a média de falsos negativos foi menor do que aquelas em ambas, pseudo-palavras e palavras de baixa freqüência. De fato, a discrepância em relação ao esperado parece tão grande que é possível dizer que o efeito desconcertante revelado na Figura 1 (à esquerda) em que palavras de alta freqüência produziram menor média de falsos negativos do que palavras de baixa freqüência, pode ser atribuído à "colaboração" dessas palavras regra de alta freqüência (já que tanto nas irregulares

quanto nas irregulares não houve diferença na média de falsos negativos entre palavras de baixa e alta freqüência). Assim, a composição do agrupamento de palavras regra de alta freqüência deve ser analisada a partir de modelos alternativos (tais como os fono-articulatórios) com vistas a identificar o que há nelas de tão especial que faz com que sejam pronunciadas com um contraste segmental mais pronunciado do que seria esperado.

A discrepância na média de falsos negativos provavelmente se deveu a uma série de pequenos desequilíbrios aparentes entre os agrupamentos de palavras de baixa e alta freqüência, especialmente nas palavras regra. Tais desequilíbrios foram de três ordens: a freqüência das palavras, o número de sílabas das palavras, e a composição fonética das palavras. Em termos de freqüência das palavras, uma comparação entre as listas de Pinheiro e de Françozo (n.d.) revela discrepâncias grandes entre a categoria de freqüência (alta x baixa) de algumas palavras na lista de Pinheiro e a freqüência real computada por Françozo. Não puderam ser encontrados dados relativos à natureza e ao tamanho e da amostra a partir da qual foram obtidas as classificações de freqüência por Pinheiro (1994). Sabe-se, no entanto, que se trata de uma lista apropriada à testagem de crianças escolares de primeira à quarta série. Já a natureza e o tamanho da amostra a partir da qual Francozo obteve o cômputo de freqüência é conhecida e bastante diferente. Trata-se de uma base de dados de cerca de 75.000 palavras diferentes obtidas a partir de uma amostra de cerca de 1.750.000 palavras, que corresponde a todo um mês da redação da Folha de São Paulo. Uma vez que os participantes do presente estudo eram jovens adultos universitários, os cômputos de freqüência de Françozo devem ser levados em consideração para uma estimativa da freqüência daquelas palavras para essa amostra de participantes, sendo talvez mais adequado a uma estimativa de freqüência de ocorrência na língua para a amostra do estudo do que as classificações de freqüência na lista de Pinheiro.

Alguns exemplos ilustrativos das discrepâncias entre as estimativas de freqüência entre Pinheiro (1994) e Françozo (n.d.) podem ser assim arrolados: na categoria das palavras de alta freqüência, Pinheiro incluiu palavras como *onça*, *pássaro*, *sílabas*, *folhas*, *papai*, *galinha* e *gato*. Buscando essas mesmas palavras na lista de Françozo obtém-se as seguintes freqüências aproximadas: 0, 0, 1, 22, 22, 43, e 61, respectivamente. Já na categoria das palavras de baixa freqüência, Pinheiro incluiu palavras como *chegada*, *mostra*, *marca*, *vejam*, *olhava*, 2210, 1022, 934, 900 e 460, respectivamente. Como seria de se esperar, fica clara a diferença entre a freqüência de ocorrência das palavras na língua para crianças e adultos. Isto parece explicar a ausência de efeito de freqüência no presente estudo. Tais considerações servem para enfatizar a necessidade de conduzir experimentos ulteriores com CronoFonos 2.0 empregando listas de palavras com freqüência apropriada aos participantes. Depois de contrabalanceada em todas as demais dimensões, a lista de freqüência de palavras Françozo deverá ser empregada com adultos, enquanto a de Pinheiro deverá ser testada com crianças no experimento proposto. No presente experimento foram empregados adultos pois o propósito era o de calibrar o instrumento computadorizado com o mínimo de variação extrínseca às variáveis; e foi empregada a lista de Pinheiro com esses adultos porque não havíamos ainda tido acesso a outras listas.

Ainda relevante à compreensão da natureza da discrepância na freqüência de falsos negativos entre palavras de alta e de baixa freqüência, é o seguinte fato:

Em termos de composição silábica, foi observado na lista de palavras de Pinheiro um maior número de sílabas nas palavras de baixa (cinco sílabas a mais) do que nas de alta freqüência. Isto é especialmente relevante uma vez que a medida aqui considerada (freqüência de falsos negativos) refere-se explicitamente à diferença entre o número de segmentos locucionais articulados pelos participantes durante a leitura em voz alta e o número de segmentos silábicos presentes na ortografia.

Finalmente, também relevante à compreensão da natureza da discrepância, é o seguinte fato: Em termos de composição fonética das palavras da lista de Pinheiro, aparentemente pode-se identificar uma maior freqüência de intervalos contrastantes no grupo de alta do que no de baixa freqüência, como se depreende das informações de Russo e Belau (1993) acerca da representação dos sons do português no audiograma. O gráfico 9 de Russo e Belau (1993) contém o registro gráfico do audiograma representando os valores acústicos médios de freqüência e intensidade dos sons da fala do português brasileiro. Tomando apenas o nível de audição (em dB) daquele registro, pode-se ordenar todos os sons do português num *continuum* de intensidade, indo desde cerca de 14 e 15 com as consoantes fricativas *v* e *f*, respectivamente, até cerca de 42 a 44, com as vogais *a* e *ə*, respectivamente. Considerando-se um modelo de contrastes de valores de nível de audição nas junções silábicas (por exemplo, a junção V1C2 na estrutura C1V1C2V2), poder-se-ia supor heuristicamente que quanto maior o contraste entre os componentes dessas junções (ou seja, a diferença entre os valores dos níveis) tanto maior a discriminabilidade dos segmentos. Tal hipótese pode ser especificamente testada usando-se CronoFonos 2.0, e este é um dos próximos passos cuja necessidade ficou clara a partir dos dados do presente estudo piloto. Uma análise preliminar na lista de Pinheiro, tomando apenas as palavras com estrutura CVCV nos agrupamentos de alta e baixa freqüência, revelou uma correlação significante entre segmentabilidade locucional (indicada por uma menor incidência de falsos negativos em CronoFonos 2.0) e os valores de uma matriz de pesos de contraste articulatório (levando em consideração os *continua* decrescente de discriminabilidade: tipo de consoante [das plosivas às fricativas e nasais], local de articulação dessas consoantes [anterior, central e posterior]; e local de articulação das vogais [anterior: *i*, *e*, ɛ; central: *a*; posterior: *ə*, *o*, *u*]). No entanto, um exame mais aprofundado faz-se necessário antes que conclusões mais fortes possam ser derivadas.

Para ilustrar as possibilidades de análise fono-articulatória e acústica, podemos tecer algumas comparações simples. Por exemplo, considerando-se que na lista de Pinheiro há 48 palavras de alta freqüência e 48 palavras de baixa freqüência, uma simples contagem revela que há mais consoantes plosivas no conjunto de palavras de alta freqüência do que naquele de baixa freqüência (51 contra 43, respectivamente). Ora, consoantes plosivas são altamente contrastantes; logo, sua maior concentração relativa no conjunto de palavras de alta freqüência tende a reduzir a freqüência relativa de falsos negativos nesse conjunto de alta freqüência. Do mesmo modo, a contagem revela que há mais consoantes nasais no conjunto de palavras de baixa freqüência do que naquele de alta (24 contra 18, respectivamente). Ora, consoantes nasais são muito pouco contrastantes; logo, sua maior concentração relativa no conjunto de palavras de baixa freqüência tende a aumentar a freqüência relativa de falsos negativos nesse conjunto de baixa freqüência. Assim, como a freqüência rela-

tiva de plosivas é maior no conjunto de palavras de alta freqüência; e a de nasais é maior no conjunto de palavras de baixa freqüência, unicamente com base nessa composição consonantal, esperar-se-ia que a freqüência de falsos negativos fosse bem maior no conjunto de baixa freqüência do que no de alta freqüência. Deste modo, pode-se dizer que a composição consonantal das palavras na lista de Pinheiro pode estar "conspirando contra" o efeito de freqüência de ocorrência das palavras na língua, tal como medida por meio de um *software* dedicado à detecção de falsos negativos com base no contraste consonantal.

CronoFonos é programado especificamente para computar com sensibilidade a freqüência de falsos negativos. Tal freqüência é afetada não apenas pelas variáveis psicolinguísticas de lexicalidade, regularidade, freqüência e comprimento às quais usualmente se atenta quando se constrói listas de palavras como a de Pinheiro (1994), como também por variáveis de natureza articulatória e fonética às quais usualmente não se atenta. Outros laboratórios de inquestionável respeitabilidade como o da Profa. Dra. Eleanora Albano na Lingüística da Unicamp com sua grande preocupação para com balanceamento fonético podem fornecer informações vitais para a elaboração de listas de palavras mais equilibradas, de modo a permitir avaliar o efeito de variáveis "elusivas" (como freqüência, ao menos no presente experimento) com maior precisão. Tal possibilidade é tornada ainda mais tangível pelo trabalho no mesmo Laboratório da equipe do Prof. Dr. Edson Françozo que vem se dedicando a elaboração de listas de freqüência de palavras. A articulação dos esforços de pesquisadores como Albano, Behlau, Françozo, Parente, Pinheiro, Russo e outros é bastante auspiciosa. Um de nossos próximos passos é precisamente a elaboração de sublistas de palavras com base nesses trabalhos para testagem via CronoFonos 2.0.

Os presentes dados deste primeiro estudo exploratório atestam a eficácia de CronoFonos 2.0 enquanto um instrumento para análise de latência, duração, e segmentação locucional durante leitura em voz alta. O presente estudo abre caminho à experimentação com variadas listas de palavras para a testagem de modelos de processamento de informação na leitura, bem como modelos fono-articulatórios. Finalmente, abrem caminho também à sua aplicação em delineamentos experimentais englobando variáveis de desenvolvimento para análise dos processos de aquisição de leitura; e eventualmente para a análise de processos subjacentes à perda de leitura subsequente a lesão cerebral (alexias adquiridas). Tais desenvolvimentos serão objetos de estudos subsequentes.

Referências Bibliográficas

Ball, E.W., & Blachman, B.A. (1991). Does phoneme awareness training in kindergarten make a difference in early word recognition and developmental spelling? *Reading Research Quarterly, 26*, 49-66.

Behlau, M.S., Pontes, P.A., Tosi, O., & Ganança, M.M. (1988a). Análise espectrográfica de formantes das vogais do português brasileiro falado em São Paulo. *Acta WHO, 7*, 67-73.

Behlau, M.S., Pontes, P.A., Tosi, O., & Ganança, M.M. (1988b). Análise perceptual acústica das vogais do português brasileiro falado em São Paulo. *Acta WHO, 7*, 74-85.

Behlau, M.S., Pontes, P.A., Tosi, O., & Ganança, M.M. (1988c). Análise do tempo de início de sonorização dos sons plosivos do português. *Acta WHO, 7*, 86-97.

Behlau, M.S., Tosi, O., & Pontes, P.A. (1985). Determinação da frequência fundamental e suas variações em altura ("jitter") e intensidade ("shimmer"), para falantes do português brasileiro. *Acta WHO, 4*, 5-9.

Byrd, D. (1996). Influences on articulatory timing in consonant sequences. *Journal of Phonetics, 24*, 209-244.

Byrne, B., & Fielding-Barnsley, R. (1989). Phonemic awareness and letter knowledge in the child's acquisition of the alphabetic principle. *Journal of Educational Psychology, 81*, 13-321.

Capovilla, F.C. (1989). *On the context of discovery in experimentation with human subjects*. Ann Arbor, MI: U.M.I.

Capovilla, F.C. (1992). Informática aplicada à educação especial. *Anais da XXI Reunião Anual de Psicologia da Sociedade Brasileira de Psicologia*, Ribeirão Preto, S.P., outubro 1991, 252-254.

Capovilla, F.C. (1993). Pesquisa e desenvolvimento de novos recursos tecnológicos para educação especial: Boas novas para pesquisadores, clínicos, professores, pais e alunos. *Em Aberto, 13*(60), 139-151. Número especial dedicado à Educação Especial. Editora-Organizadora: E.M.L.S. de Alencar., Brasília, D.F.

Capovilla, F.C. (1994). Sistemas ampliam a comunicação de deficientes. *Diálogo Médico, 20*(1), 24-27.

Capovilla, F.C. (1994). Informática em educação especial: comunicação na ausência de fala. Texto completo publicado em *Proceedings of the XVII International School Psychology Colloquium e Anais do II Congresso Nacional de Psicologia Escolar*, Campinas, S.P., julho, 155-158.

Capovilla, F.C. (no prelo). *Novos paradigmas em clínica e educação: Usando o computador numa abordagem neuropsicolinguística*. São Paulo, S.P.: Robe Editorial.

Capovilla, F.C. (1996). Sistemas especialistas de multimídia em educação especial. Em L.R.O.P. Nunes (Ed.). *Prevenção e intervenção em educação especial*. Rio de Janeiro, R.J.: Série Coletâneas da ANPEPP, vol 14. Rio de Janeiro, R.J.: Associação Nacional de Pesquisa e Pós-Graduação em Psicologia, 124-150.

Capovilla, F.C., Capovilla, A.G.S., & Colorni, E.M. (no prelo). Leitura, ditado, e manipulação fonêmica em função de variáveis psicolinguísticas em escolares de terceira a quinta séries com dificuldades de aprendizagem. *Revista Brasileira de Educação Especial*.

Capovilla, F.C., Capovilla, A.G.S., Macedo, E.C., Costa, C.E., & Duduchi, M. (no prelo). Manipulação de envolvimento de ego via para-instruções experimentais: efeitos sobre estados de ânimo e desempenho edutivo em resolução de problemas. *Psicologia USP*, São Paulo, S.P.

Capovilla, F.C., César, O.P., Wolf, R.L., & Seabra, A.G. (1993). Equ-Aritmética: Programa para análise de raciocínio aritmético em escolares de primeiro grau.

Anais da I Jornada USP-SUCESU-SP de Informática e Telecomunicações, São Paulo, S.P., julho, 501-508.

Capovilla, F.C., Colorni, E.M.R., Nico, A.M., & Capovilla, A.G.S. (1995a). Leitura em voz alta, tomada de ditado, manipulação fonêmica, e relações entre elas: Efeito de características de palavras (frequência, regularidade, lexicalidade) e de nível de escolaridade. In B.P. Damasceno & Coudry, M.I.H. (Orgs.). *Temas em Neuropsicologia e Neurolinguística, 4*, 157-169.

Capovilla, F.C., Gonçalves, M.J., Macedo, E.C., Duduchi, M., & Capovilla, A.G.S. (1996). Evidence of verbal processes in message encoding by cerebral-palsied using a picto-ideographic AAC system. *Proceeding of the VII Biennial Conference of the International Society for Augmentative and Alternative Communication*, Vancouver, B.C., Canada, 450-451.

Capovilla, F.C., Haydu, V.B., Costa, C.E., Luzia, J.C., Andrade, M.P., Silva, L.S., Macedo, E.C., Duduchi, M., & Capovilla, A.G.S. (1995). Edução de regras em Nomos v3: Efeito de condições de resolução silenciosa, verbalizada, e com interferência, sob diferentes graus de dificuldade de edução. *Torre de Babel: Pesquisa e Reflexões em Psicologia, 2*, 30-38.

Capovilla, F.C., & Macedo, E.C. (1994). Ferramentas de informática em pesquisa e prática psicopedagógica. Texto completo publicado em *Proceedings of the XVII International School Psychology Colloquium* e *Anais do II Congresso Nacional de Psicologia Escolar*, Campinas, S.P., julho, 121-125.

Capovilla, F.C., Macedo, E.C., Duduchi, M., Capovilla, A., & Gonçalves, M.J. (1996). "Brincar de Ler": O computador no diagnóstico diferencial das dislexias. *O Mundo da Saúde, 20*(2), 87-89.

Capovilla, F.C., Macedo, E.C., Duduchi, M., Capovilla, A.G.S., Raphael, W.D., & Guedes, M. (1996). Ultr*AAC*tive: Computerized multimedia expert AAC system. *Proceedings of the VII Biennial Conference of the International Society for Augmentative and Alternative Communication*. Vancouver, B.C., Canada, August, 1996, 452-453.

Capovilla, F.C., Macedo, E.C., Duduchi, M., Gonçalves, M.J., & Capovilla, A.G.S. (1996). Home use of a computerized pictographic-syllabic-vocalic AAC system in cerebral palsy: preliminary data. *Proceedings of the VII Biennial Conference of the International Society for Augmentative and Alternative Communication*. Vancouver, B.C., Canada, August, 1996, 454-455.

Capovilla, F.C,. Macedo, E.C., Duduchi, M., Raphael, W.D., Guedes, M., Capovilla, A.G.S., & Gonçalves, M.J. (1996). Avaliação computadorizada de vocabulário e compreensão auditiva em crianças falantes ou não. *O Mundo da Saúde, 20*(1), 421-424.

Capovilla, F.C,. Macedo, E.C., Duduchi, M., Thiers, V.O., Capovilla, A.G.S., & Gonçalves, M.J. (1995). Como selecionar o melhor sistema de comunicação para seu paciente com deficit de fala? *O Mundo da Saúde, 19*(10), 350-352.

Capovilla, F.C., Macedo, E.C., Duduchi, M., & Guedes, M. (1996). "Synthesized or digitized speech? That is the question." NoteVox: Portable computerized AAC system with digitized voice for anartria, cerebral palsy and amyotrophic lateral

sclerosis. *Proceedings of the VII Biennial Conference of the International Society for Augmentative and Alternative Communication*. Vancouver, B.C., Canada, August, 1996, 456-457.

Capovilla, F.C., Macedo, E.C., Duduchi, M., Raphael, W.D., Capovilla, A.G.S., & Guedes, M. (1996). Computerized tools for assessing scholastic progress in AAC users with severe motor impairments. *Proceedings of the VII Biennial Conference of the International Society for Augmentative and Alternative Communication*. Vancouver, B.C., Canada, 55-56.

Capovilla, F.C., Macedo, E.C., Feitosa, M.D., & Seabra, A.G. (1993). Mestre Sonorizado, Escriba, Emulador de tela sensível ao toque: Soluções de baixa tecnologia para problemas em alta (diagnóstico e terapia em psicopedagogia). *Anais da I Jornada USP-SUCESU-SP de Informática e Telecomunicações*, São Paulo, S.P., julho, 483-492.

Capovilla, F.C., Macedo, E.C., Feitosa, M.D., & Seabra, A.G. (1993). ImagoVox: Portavoz eletrônico para pacientes neurológicos. *Anais da I Jornada USP-SUCESU-SP de Informática e Telecomunicações*, São Paulo, S.P., 443-448.

Capovilla, F.C., Macedo, E.C., Gonçalves, M.J., Capovilla, A.G.S., Raphael, W.D., Colorni, E., & Duduchi, M. (1995). Computer systems for assessing reading, writing, and phonologic segmentation abilities. *Proceedings of the European Conference on the Advancement of Rehabilitation Technology*. Lisbon, Portugal, 86-88.

Capovilla, F.C., Macedo, E.C., Raphael, W.D., Capovilla, A.G.S., Gonçalves, M.J., Duduchi, M., & Guedes, M. (1995). Multimedia expert systems for cognitive evaluation of AAC system users in special education. *Annals of the Third ECART European Conference on the Advancement of Rehabilitation Technology*, Lisbon, Portugal, 89-91.

Capovilla, F.C., Macedo, E.C., Raphael, W.D., Duduchi, M., Moreira, M.A.C., Gonçalves, M.J., & Capovilla, A.G.S. (1995). Multimedia expert systems for communication and education of the hearing impaired. *Annals of the Third ECART European Conference on the Advancement of Rehabilitation Technology*, Lisbon, Portugal, 83-85.

Capovilla, F.C., Macedo, E.C., Seabra, A.G., Feitosa, M.D., & Thiers, V.O. (1994). Sistemas computadorizados para surdo-mudos baseados em língua de sinais: Comunicação via Logofone, e ensino via Logofone Tutor. *Anais da II Jornada USP-SUCESU-SP de Informática e Telecomunicações*, São Paulo, S.P., junho, 363-372.

Capovilla, F.C., Nunes, L.R.O.P., Araújo, I., Nunes, D., Nogueira, D., Bernat, A.B., Ribeiro, A., Carvalho, V., & Capovilla, A.G.S. (no prelo, b). Normatização fluminense do Teste de Vocabulário por Imagens Peabody. *Revista Brasileira de Educação Especial*.

Capovilla, F.C., Raphael, W.D., Capovilla, A.G.S., Guedes, M., Costa, C.E., Macedo, E.C., Duduchi, M., Aligieri, S., Santos, A., Viana, A.L.A.G., Fuso, S.F., & Gonçalves, M.J. (1996). Sistema de multimídia para comunicação surdo-surdo e surdo-ouvinte em línguas brasileira e americana de sinais via redes de computador. *O Mundo da Saúde, 20*(3), 110-114.

Cluff, M. & Luce, P. (1990). Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation. Journal of Experimental Psychology: *Human Perception & Performance, 16*, 551-563.

Duduchi, M., Macedo, E.C., Capovilla, F.C., Capovilla, A.G.S., Gonçalves, M.J., & Thiers, V.O. (1995). ImagoAnaVox: Sistema computadorizado de comunicação alternativa. *Annals of the Third ECART European Conference on the Advancement of Rehabilitation Technology*, Lisboa, Portugal, 415.

Duduchi, M., Macedo, E.C. Soria, R., & Capovilla, F.C. (1995). CronoFonos: Sistema para registro em milisegundos de latência e duração de leitura em voz alta em computadores convencionais para estudos em teoria de processamento de informação. *Resumos da XXV Reunião Anual de Psicologia da Sociedade Brasileira de Psicologia,* Ribeirão Preto, S.P., 228.

Dunn, L.M., Padilla, E.R., Lugo, D.E., & Dunn, L.M. (1986). *Manual del examinador para el Test de Vocabulario en Imágenes Peabody*. American Guidance Service, Circle Pines, MN.

Ellis, A., & Young, A.W. (1988). *Human cognitive neuropsychology*. London, U.K.: Lawrence Erlbaum.

Eysenck, M.W., & Keane, M.T. *Psicologia cognitiva: Um manual introdutório*. Porto Alegre, RS: Editoria Artes Médicas Sul.

Feitosa, M.D., Macedo, E.C., Capovilla, F.C., Seabra, A.G., & Thiers, V.O. (1994). Sistemas computadorizados de comunicação e de ensino para paralisia cerebral baseados na linguagem Bliss. *Anais da II Jornada USP-SUCESU-SP de Informática e Telecomunicações*, São Paulo, S.P., junho, 343-352.

Ferreira, F., Henderson, J., Anes, M.D., Weeks, P.A., & McFarlane, D.K. (1996). Effects of lexical frequency and syntactic complexity in spoken-language comprehension: Evidence from the auditory moving window technique. *Journal of Experimental Psychology: Learning, Memory & Cognition, 22*, 325-335.

Françozo, E. (não publicado). *Lista de frequência de palavras*. Universidade Estadual de Campinas, Campinas, S.P.

Frith, U. (1985). Beneath the surface of developmental dyslexia. In K. Patterson, J. Marshall, & M. Coltheart (Eds.). *Surface dyslexia: neuropsychological and cognitive studies of phonological reading*. London: Erlbaum.

Galaburda, A.M. (1989). (Org.). *From Reading to Neurons*. Cambridge, Massachusetts: MIT Press.

Gielow, I. (1993). *Análise espectrográfrica da zona de transição dos formantes das vogais subsequentes aos sons plosivos do português brasileiro*. Monografia de especialização. Escola Paulista de Medicina.

Gonçalves, M.J., Capovilla, F.C., & Macedo, E.C. (no prelo). A fonoaudiologia na era da informática e seu encontro com a comunicação alternativa e facilitadora. Em C. César e M. Lagrota (Orgs.). *Tópicos em Fonoaudiologia*. São Paulo, S.P.: Editora Lovise.

Gonçalves, M.J., Capovilla, F.C., Macedo, E.C., Feitosa, M.D., & Seabra, A.G. (1995). Quando falar não é possível: Uma alternativa. *Cadernos de Estudos Linguísticos, 29*, 49-56.

Gonçalves, M.J., Macedo, E.C., Duduchi, M., Capovilla, A.G.S., Thiers, V.O., & Capovilla, F.C. (1995). Comunicação computadorizada a serviço de saúde e qualidade de vida. *O Mundo da Saúde, 19*(4), 145-148.

Gonçalves, M.J., Macedo, E.C., Duduchi, M., & Capovilla, F.C. (1995). Computerized Pictogram Ideogram Communication System for cerebral palsy, preliminary data. *Annals of the Third ECART European Conference on the Advancement of Rehabilitation Technology,* Lisbon, Portugal, 92-94.

Hahn, M. & Park, C.K. (1992). An improved speech detection algorithm for isolated Korean utterances. *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Proc.,* 525-528.

Haydu, V.B, Capovilla, F.C., Jacob, A.N., Pimentel, N.S., Guilherme, R.L., Macedo, E.C., Duduchi, M., & Capovilla, A.G.S. (1995). Resolução de problemas: Efeito do grau de dificuldade de edução durante o learning set. *Torre de Babel: Pesquisa e Reflexões em Psicologia, 2*, 21-29.

Herzel, H., Berry, D., Titze, I.R., & Saleh, M. (1994). Analysis of vocal disorders with methods from nonlinear dynamics. *Journal of Speech and Hearing Research, 37*, 1008-1019.

Hohn, W.E., & Ehri, L.C. (1983). Do alphabet letters help prereaders acquire phonemic segmental skill? *Journal of Educational Psychology, 75*, 752-762.

Hosogi, M.L. & Parente, M.A.P. (1995). As dislexias adquiridas com utilização da via perilexical: manifestações das dislexias de superfície. Em B.P. Damasceno e M.I.H. Coudry. *Temas em Neuropsicologia e Neurolinguística. Série de Neuropsicologia, 4*, Sociedade Brasileira de Neuropsicologia, 174-179.

Juel, C., Griffith, P.L., & Gough, P.B. (1986). Acquisition of literacy: A longitudinal study of children in first and second grade. *Journal of Educational Psychology, 78*, 243-255.

Lamel, L.F. et al. (1981). An improved endpoint detector for isolated word recognition. *IEEE Trans. Acoust., Speech, and Signal Processing, 29*(4), 777-785.

Larar, J.N. (1985). *Towards speaker-independent isolated word recognition for large lexicons: A two-channel, two pass approach*, Unpublished Doctoral Dissertation, University of Florida.

Lecours, A.R., Delgado, A.P., & Pimenta, M.A.M. (1993). Distúrbios adquiridos da leitura e da escrita. Em Mansur, L. e Rodrigues, N. *Temas em Neurolinguística, Série de Neuropsicologia, 3*, São Paulo, S.P.: Sociedade Brasileira de Neuropsicologia, 31-44.

Lee, H.S. & Hahn. M. (1994). *Development of an algorithm for the determination of endpoints in real time*. Unpublished manuscript. Electronics and Telecommunications Research Institute, Korea.

Lemle, M. (1991). *Guia teórico do alfabetizador*. 6a. ed., São Paulo, S.P. Editora Ática.

Lundberg, I., Frost, J., & Petersen, D.P. (1988). Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading Research Quarterly, 23*, 264-284.

Macedo, E.C., Capovilla, F.C., Gonçalves, M.J., Seabra, A.G., Thiers, V.O., & Feitosa, M.D. (1994). Adaptando um sistema computadorizado pictográfico para comunicação em paralisia cerebral tetra-espástica. *Anais da II Jornada USP-SUCESU-SP de Informática e Telecomunicações*, São Paulo, S.P., junho, 353-362.

Macedo, E.C., Duduchi, M., Soria, R., & Capovilla, F.C. (não publicado). *CronoFonos 2.0: implementação do algoritmo de Lee e Hahn ajustado para cômputo em tempo real de latência e duração locucional e de frequência e duração de segmentos locucionais em Português*. manuscrito não publicado, Universidade de São Paulo, São Paulo, S.P.

Mann, V., & Brady, S. (1988). Reading disability: The role of language deficiencies. *Journal of Consulting and Clinical Psychology, 56*, 811-816.

Marshall, J. (1989). The description and interpretation of acquired and developmental reading disorders. In A.M. Galaburda (Org.). *From Reading to Neurons*. Cambridge, Massachusetts: MIT Press.

Metz-Lutz, M.N. (1993). Neuropsicolinguística e reeducação da afasia. Em L.L. Mansur & N. Rodrigues (Eds.). *Temas em Neurolinguística, 2*, Sociedade Brasileira de Neuropsicologia. São Paulo, S.P.; Sociedade Brasileira de Neuropsicologia, 107-115.

Milenkovic, P. (1987). Least mean square measures of voice perturbation. *Journal of Speech and Hearing Research, 30*, 529-538.

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review, 76*, 165-178.

Morton, J. (1979). Facilitation in word recognition: Experiments causing change in the Logogen Model. Em P.A. Kohlers, M.E. Wrolstand, e H. Bouma (Eds.). *Processing of visible language, 1*, New York, N.Y., 259-268.

Morton, J. (1989). An information-processing account of reading acquisition. In A.M. Galaburda (Org.). *From Reading to Neurons*. Cambridge, Massachusetts: MIT Press.

Neuburg, E.P. (1979). Automatic thresholding for voicing detection algorithms. Proc. *IEEE Int. Conf. on Acoust., Speech, and Signal Proc.*, 756-759.

O'Connor, R.E., Jenkins, J.R., Leicester, N., & Slocum, T.A. (1993). Teaching phonological awareness to young children with learning disabilities. *Exceptional Children, 59*, 532-546.

Parente, M.A.P. (1995). O enfoque cognitivo na Avaliação das dislexias adquiridas e o sistema ortográfico do português. Em B.P. Damasceno e M.I.H. Coudry. *Temas em Neuropsicologia e Neurolinguística. Série de Neuropsicologia, 4*, Sociedade Brasileira de Neuropsicologia, 169-173.

Perfetti, C.A., Beck, I., Ball, L.C., & Hughes, C. (1987). Phonemic knowledge and learning to read are reciprocal: A longitudinal study of first grade children. *Merrill-Palmer Quarterly, 33*, 283-319.

Pinheiro, A.M.V. (1994). *Leitura e escrita: Uma abordagem cognitiva*. Campinas, S.P.: Editorial Psy II.

Pratt, A., & Brady, S. (1988). Relation of phonological awareness to reading disability in children and adults. *Journal of Educational Psychology, 80*, 319-323.

Rabiner, L.R. & Sambur, M.R. (1975). An algorithm determining de endpoints of isolated utterance. *Bell Syst. Tech. J., 54*(2). 297-315.

Rabiner, L.R. & Schafer, L.W. (1978) *Digital Processing of Speech Signals*. England Cliffs, NJ., Prentice-Hall, 1978.

Rohl, M., & Tunmer, W.E. (1988). Phonemic segmentation skill and spelling acquisition. *Applied Psycholinguistics, 9*, 335-350.

Rosenthal, R. & Rosnow, R.L. (1984). *Essentials of behavioral research: Methods and data analysis*. New York, N.Y.: McGraw-Hill.

Russo, I. & Behlau, M.S. (1993). *Percepção da fala: Análise acústica*. São Paulo, S.P.: Editora Lovise.

Santos, M.T.M. (1987). *Uma análise espectrográfica dos sons fricativos surdos e sonoros do português brasileiro*. Monografia de especialização. Escola Paulista de Medicina.

Seidenberg, M.S., & McClelland, J.L. (1989). Visual word recognition and pronunciation: A computational model of acquisition, skilled performance, and dyslexia. In A.M. Galaburda (Ed.), *From reading to neurons*. Cambridge, MA: MIT Press.

Serón, X. (1993). A reeducação neuropsicológica: As abordagens cognitivista e pragmática. Em L.L. Mansur & N. RoDrigues (Eds.). *Temas em Neurolinguística., 2*, Sociedade Brasileira de Neuropsicologia. São Paulo, S.P.; Sociedade Brasileira de Neuropsicologia, 122-144.

Seymour, P.H. & Pinheiro, A.M. (1995). Aporte cognitivista à avaliação do desenvolvimento da leitura. Em B.P. Damasceno e M.I.H. Coudry. *Temas em Neuropsicologia e Neurolinguística. Série de Neuropsicologia, 4*, Sociedade Brasileira de Neuropsicologia, 142-148.

Shallice, T. (1990). *From neuropsychology to mental structure*. Cambridge, U.K.: Cambridge University Press.

Soria, R. (não publicado). *Adaptação e correção do algoritmo de Lee e Hahn para cômputo em tempo real de início e término de locuções*. Manuscrito não publicado, Universidade de São Paulo, São Paulo, S.P.

Stanovich, K.E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360-407.

Stenberger, J.G. & Treiman, R. (1986). The internal structure of word-initial consonant clusters. *Journal of Memory and Language, 25*, 163-180.

Sulter, A.M., Wit, H.P., Schutte, H.K., & Miller, D.G. (1994). A structures approach to voice range profile (phonetogram) analysis. *Journal of Speech and Hearing Research, 7*, 1076-1085.

Telage, K. M. (1980). A computerized place-manner distinctive feature program for articulation analysis. *Journal of Speech and Hearing Disorders, 45*, 481-494.

Thiers, V.O., & Capovilla, F.C. (1996). Alternative communication in cerebral palsy: evaluation of variables that control the search for Blissymbols on communication boards. *Proceedings of the VII Biennial Conference of the International Society for Augmentative and Alternative Communication*. Vancouver, B.C., Canada, August, 60-61.

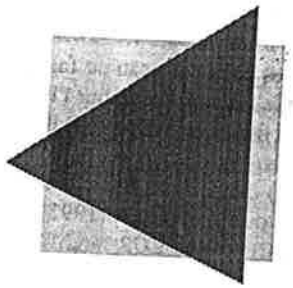
Thiers, V.O., Capovilla, F.C., Macedo, E.C., Feitosa, M.D., & Seabra, A.G. (1994). Aplicação do *software* Sonda para análise diferencial de iconicidade em sistemas de comunicação para pacientes neurológicos. *Anais da II Jornada USP-SUCESU-SP de Informática e Telecomunicações*, São Paulo, S.P., junho, 373-382.

Throneburg, R.N. & Yairi, E. (1994). Temporal dynamics of repetitions during the early stage of childhood stuttering: An acoustic study. *Journal of Speech and Hearing Research, 37*, 1067-1075.

Warrick, N., Rubin, H., & Rowe-Walsh, S. (1993). Phoneme awareness in language-delayed children: comparative studies and intervention. *Annals of Dyslexia, 43*, 153-173.

Wilpon, J.G. & Rabiner, L.R (1987). Application of hidden Markov models to automatic speech endpoint detection. *Computer Speech and Language, 2*, 321-341.

Yairi, E., & Lewis, B. (1984). Disfluencies at the onset of stuttering. *Journal of Speech and Hearing Research, 27*, 154-159.



ANEXO 1

O ALGORITMO PARA A DETERMINAÇÃO DE *ENDPOINTS* EM TEMPO REAL EMPREGADO NO *SOFTWARE* CRONOFONOS 2.0.

É muito importante determinar precisamente o início e o fim de uma locução desconhecida quando se deseja realizar o reconhecimento de uma locução com alto desempenho por meio de um sistema de análise de fala. Lee e Hahn (1994) descreveram um algoritmo para determinação de *endpoints* desenvolvido para PC 486. O *corpus* ou base de dados para o algoritmo era composto de 37 palavras. A sala de teste era um escritório comum, com ruídos de fundo como o de digitação em teclado, passos, e ventiladores. Todos os sinais eram filtrados e depois digitalizados em 10 Khz com resolução de 16 *bits*. Os dados para a tomada de decisão eram *energia* e *taxa de cruzamento de zeros* (ou nível de cruzamento) *ambos calculados para cada quadro de 10 ms*. A avaliação do desempenho do algoritmo com cinco locutores mostrou erro médio de 3.1 ms para detecção de início de locução e cerca de 6.4 ms para detecção do término. Comparando tais resultados com os da literatura, Lee e Hahn concluíram que o seu pode ser considerado como um dos melhores algoritmos de tempo real, sendo ainda relativamente simples.

Distinguir precisamente uma vocalização de um ruído de fundo é muito importante em análise de fala e reconhecimento de palavra. Em sistemas de reconhecimento de palavras isoladas (IWR *isolated word recognition*) a identificação precisa de fala se dá a partir da identificação dos *endpoints*. O consumo de recursos computacionais pode ser significantemente reduzido se a identificação for realizada ao mesmo tempo em que os dados externos são descarregados para armazenamento eficiente. Consequentemente, a qualidade do detector de fala é diretamente afetada pelo desempenho de sistemas de IWR, pois eles freqüentemente usam informações baseadas em *endpoints* com técnicas de *dynamic time warping* (DTW) para verificar o sucesso do emparelhamento entre uma locução desconhecida e um modelo previamente armazenado. Mesmo para os casos de um sistema IWR baseados no modelo de Markov (HMM *hidden Markov model*) ou baseados em redes neurais seria uma grande vantagem se informações precisas pudessem ser fornecidas.

O problema de detecção de fala usualmente parece muito simples mas é não trivial, exceto nos casos de ambientes com alta razão entre sinal e ruído (*S/N signal-to-noise*). Mesmo quando uma alta razão S/N é garantida, a energia do mais baixo nível de som de fala facilmente excede a energia do som de fundo, e este pode ser um limiar estável para a produção de resultados satisfatórios. Para Rabiner e Schafer

Lundberg, I., Frost, J., & Petersen, D.P. (1988). Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading Research Quarterly, 23*, 264-284.

Macedo, E.C., Capovilla, F.C., Gonçalves, M.J., Seabra, A.G., Thiers, V.O., & Feitosa, M.D. (1994). Adaptando um sistema computadorizado pictográfico para comunicação em paralisia cerebral tetra-espástica. *Anais da II Jornada USP-SUCESU-SP de Informática e Telecomunicações*, São Paulo, S.P., junho, 353-362.

Macedo, E.C., Duduchi, M., Soria, R., & Capovilla, F.C. (não publicado). *CronoFonos 2.0: implementação do algoritmo de Lee e Hahn ajustado para cômputo em tempo real de latência e duração locucional e de frequência e duração de segmentos locucionais em Português*. manuscrito não publicado, Universidade de São Paulo, São Paulo, S.P.

Mann, V., & Brady, S. (1988). Reading disability: The role of language deficiencies. *Journal of Consulting and Clinical Psychology, 56*, 811-816.

Marshall, J. (1989). The description and interpretation of acquired and developmental reading disorders. In A.M. Galaburda (Org.). *From Reading to Neurons*. Cambridge, Massachusetts: MIT Press.

Metz-Lutz, M.N. (1993). Neuropsicolinguística e reeducação da afasia. Em L.L. Mansur & N. Rodrigues (Eds.). *Temas em Neurolinguística, 2*, Sociedade Brasileira de Neuropsicologia. São Paulo, S.P.; Sociedade Brasileira de Neuropsicologia, 107-115.

Milenkovic, P. (1987). Least mean square measures of voice perturbation. *Journal of Speech and Hearing Research, 30*, 529-538.

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review, 76*, 165-178.

Morton, J. (1979). Facilitation in word recognition: Experiments causing change in the Logogen Model. Em P.A. Kohlers, M.E. Wrolstand, e H. Bouma (Eds.). *Processing of visible language, 1*, New York, N.Y., 259-268.

Morton, J. (1989). An information-processing account of reading acquisition. In A.M. Galaburda (Org.). *From Reading to Neurons*. Cambridge, Massachusetts: MIT Press.

Neuburg, E.P. (1979). Automatic thresholding for voicing detection algorithms. Proc. *IEEE Int. Conf. on Acoust., Speech, and Signal Proc.*, 756-759.

O'Connor, R.E., Jenkins, J.R., Leicester, N., & Slocum, T.A. (1993). Teaching phonological awareness to young children with learning disabilities. *Exceptional Children, 59*, 532-546.

Parente, M.A.P. (1995). O enfoque cognitivo na Avaliação das dislexias adquiridas e o sistema ortográfico do português. Em B.P. Damasceno e M.I.H. Coudry. *Temas em Neuropsicologia e Neurolinguística. Série de Neuropsicologia, 4*, Sociedade Brasileira de Neuropsicologia, 169-173.

Perfetti, C.A., Beck, I., Ball, L.C., & Hughes, C. (1987). Phonemic knowledge and learning to read are reciprocal: A longitudinal study of first grade children. *Merrill-Palmer Quarterly, 33*, 283-319.

Pinheiro, A.M.V. (1994). *Leitura e escrita: Uma abordagem cognitiva*. Campinas, S.P.: Editorial Psy II.

Pratt, A., & Brady, S. (1988). Relation of phonological awareness to reading disability in children and adults. *Journal of Educational Psychology, 80*, 319-323.

Rabiner, L.R. & Sambur, M.R. (1975). An algorithm determining de endpoints of isolated utterance. *Bell Syst. Tech. J., 54*(2). 297-315.

Rabiner, L.R. & Schafer, L.W. (1978) *Digital Processing of Speech Signals*. England Cliffs, NJ., Prentice-Hall, 1978.

Rohl, M., & Tunmer, W.E. (1988). Phonemic segmentation skill and spelling acquisition. *Applied Psycholinguistics, 9*, 335-350.

Rosenthal, R. & Rosnow, R.L. (1984). *Essentials of behavioral research: Methods and data analysis*. New York, N.Y.: McGraw-Hill.

Russo, I. & Behlau, M.S. (1993). *Percepção da fala: Análise acústica*. São Paulo, S.P.: Editora Lovise.

Santos, M.T.M. (1987). *Uma análise espectrográfica dos sons fricativos surdos e sonoros do português brasileiro*. Monografia de especialização. Escola Paulista de Medicina.

Seidenberg, M.S., & McClelland, J.L. (1989). Visual word recognition and pronunciation: A computational model of acquisition, skilled performance, and dyslexia. In A.M. Galaburda (Ed.), *From reading to neurons*. Cambridge, MA: MIT Press.

Serón, X. (1993). A reeducação neuropsicológica: As abordagens cognitivista e pragmática. Em L.L. Mansur & N. RoDrigues (Eds.). *Temas em Neurolinguística., 2*, Sociedade Brasileira de Neuropsicologia. São Paulo, S.P.; Sociedade Brasileira de Neuropsicologia, 122-144.

Seymour, P.H. & Pinheiro, A.M. (1995). Aporte cognitivista à avaliação do desenvolvimento da leitura. Em B.P. Damasceno e M.I.H. Coudry. *Temas em Neuropsicologia e Neurolinguística. Série de Neuropsicologia, 4*, Sociedade Brasileira de Neuropsicologia, 142-148.

Shallice, T. (1990). *From neuropsychology to mental structure*. Cambridge, U.K.: Cambridge University Press.

Soria, R. (não publicado). *Adaptação e correção do algoritmo de Lee e Hahn para cômputo em tempo real de início e término de locuções*. Manuscrito não publicado, Universidade de São Paulo, São Paulo, S.P.

Stanovich, K.E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360-407.

Stenberger, J.G. & Treiman, R. (1986). The internal structure of word-initial consonant clusters. *Journal of Memory and Language, 25*, 163-180.

Sulter, A.M., Wit, H.P., Schutte, H.K., & Miller, D.G. (1994). A structures approach to voice range profile (phonetogram) analysis. *Journal of Speech and Hearing Research, 7*, 1076-1085.

Telage, K. M. (1980). A computerized place-manner distinctive feature program for articulation analysis. *Journal of Speech and Hearing Disorders, 45*, 481-494.

Thiers, V.O., & Capovilla, F.C. (1996). Alternative communication in cerebral palsy: evaluation of variables that control the search for Blissymbols on communication boards. *Proceedings of the VII Biennial Conference of the International Society for Augmentative and Alternative Communication*. Vancouver, B.C., Canada, August, 60-61.

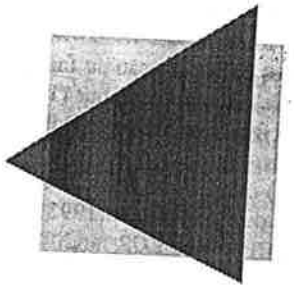
Thiers, V.O., Capovilla, F.C., Macedo, E.C., Feitosa, M.D., & Seabra, A.G. (1994). Aplicação do *software* Sonda para análise diferencial de iconicidade em sistemas de comunicação para pacientes neurológicos. *Anais da II Jornada USP-SUCESU-SP de Informática e Telecomunicações*, São Paulo, S.P., junho, 373-382.

Throneburg, R.N. & Yairi, E. (1994). Temporal dynamics of repetitions during the early stage of childhood stuttering: An acoustic study. *Journal of Speech and Hearing Research, 37*, 1067-1075.

Warrick, N., Rubin, H., & Rowe-Walsh, S. (1993). Phoneme awareness in language-delayed children: comparative studies and intervention. *Annals of Dyslexia, 43*, 153-173.

Wilpon, J.G. & Rabiner, L.R (1987). Application of hidden Markov models to automatic speech endpoint detection. *Computer Speech and Language, 2*, 321-341.

Yairi, E., & Lewis, B. (1984). Disfluencies at the onset of stuttering. *Journal of Speech and Hearing Research, 27*, 154-159.



ANEXO 1

O ALGORITMO PARA A DETERMINAÇÃO DE *ENDPOINTS* EM TEMPO REAL EMPREGADO NO *SOFTWARE* CRONOFONOS 2.0.

É muito importante determinar precisamente o início e o fim de uma locução desconhecida quando se deseja realizar o reconhecimento de uma locução com alto desempenho por meio de um sistema de análise de fala. Lee e Hahn (1994) descreveram um algoritmo para determinação de *endpoints* desenvolvido para PC 486. O *corpus* ou base de dados para o algoritmo era composto de 37 palavras. A sala de teste era um escritório comum, com ruídos de fundo como o de digitação em teclado, passos, e ventiladores. Todos os sinais eram filtrados e depois digitalizados em 10 Khz com resolução de 16 *bits*. Os dados para a tomada de decisão eram *energia* e *taxa de cruzamento de zeros* (ou nível de cruzamento) *ambos calculados para cada quadro de 10 ms*. A avaliação do desempenho do algoritmo com cinco locutores mostrou erro médio de 3.1 ms para detecção de início de locução e cerca de 6.4 ms para detecção do término. Comparando tais resultados com os da literatura, Lee e Hahn concluíram que o seu pode ser considerado como um dos melhores algoritmos de tempo real, sendo ainda relativamente simples.

Distinguir precisamente uma vocalização de um ruído de fundo é muito importante em análise de fala e reconhecimento de palavra. Em sistemas de reconhecimento de palavras isoladas (IWR *isolated word recognition*) a identificação precisa de fala se dá a partir da identificação dos *endpoints*. O consumo de recursos computacionais pode ser significantemente reduzido se a identificação for realizada ao mesmo tempo em que os dados externos são descarregados para armazenamento eficiente. Consequentemente, a qualidade de detector de fala é diretamente afetada pelo desempenho de sistemas de IWR, pois eles freqüentemente usam informações baseadas em *endpoints* com técnicas de *dynamic time warping* (DTW) para verificar o sucesso do emparelhamento entre uma locução desconhecida e um modelo previamente armazenado. Mesmo para os casos de um sistema IWR baseados no modelo de Markov (HMM *hidden Markov model*) ou baseados em redes neurais seria uma grande vantagem se informações precisas pudessem ser fornecidas.

O problema de detecção de fala usualmente parece muito simples mas é não trivial, exceto nos casos de ambientes com alta razão entre sinal e ruído (*S/N signal-to-noise*). Mesmo quando uma alta razão S/N é garantida, a energia do mais baixo nível de som de fala facilmente excede a energia do som de fundo, e este pode ser um limiar estável para a produção de resultados satisfatórios. Para Rabiner e Schafer

(1978) uma falha no algoritmo de detecção de fala é possível, mesmo quando a razão S/N ultrapassa 30 dB. Contudo, isto não tende a ocorrer em situações reais.

Alguns exemplos importantes de algoritmos para detecção de fala relatados na literatura de reconhecimento de fala inclui o de Rabiner e Sambur (1975) que usa a energia da fala e a taxa de cruzamento de zero da fala; o de Wilpon e Rabiner (1987) que adota o modelo de Markov, o de Lamel e cols. (1981) que usa pulsos de energia com um equalizador de nível; o de Larar (1985) que usa os dois canais de informações de fala e eletroglotográficas; e finalmente o de Hahn e Park (1992) que baseia-se na energia, na taxa de cruzamento de zero (ZCR), e no ZCR modificado calculado em sinais originais e predeterminados. Finalmente há ainda o algoritmo de Neuburg (1979) que foi baseado principalmente na medida de energia de baixa frequência, e tentou reduzir erros de detecção de fala que ocorriam em intervalos de locução. No entanto, nenhum de tais algoritmos é executado em tempo real. O desenvolvimento de algoritmos de determinação de *endpoints* em tempo real é extremamente útil para a implementação de sistemas para fins de pesquisa adotando técnicas de reconhecimento de palavras isoladas.

Lee e Hahn (1994) apresentaram um algoritmo simples para detecção de fala em tempo real que foi avaliado em locuções de palavras coreanas isoladas. Ele usa duas características baseadas nos quadros: energia e nível da taxa de cruzamento (LCR *level crossing rate*). Usando os valores de limiares calculados para estas duas características para a determinação de parâmetro, cada quadro com dados de entrada de fala é caracterizado como fala ou silêncio. O algoritmo foi avaliado e seus resultados foram comparados com os de Rabiner e Sambur (1975) e de Hahn e Park (1992), todos os três implementados em PC 486. Nesse experimento de Lee e Hahn (1994) a base de dados ou *corpus* era composta de 37 palavras coreanas isoladas. Cinco locutores homens pronunciavam a lista completa três vezes diretamente ao microfone em uma sala comum, com ruído de fundo. O número total de locuções desse experimento era de 555 (37 X 5 X 3). Todos os sinais eram filtrados (70 Hz - 4,5 kHz) e *sampled* em 10 kHz com resolução de 16 bits. Das 555 locuções, 111 (a lista completa falada apenas uma vez por três locutores) eram usadas como dados de treino e as outras 444 como dados de teste.

Como se sabe, é quase impossível construir um algoritmo universal para a determinação de *endpoints* com valores fixos de limiar, devido ao problema de variação do ruído de fundo em situações reais. Por esse motivo, no algoritmo de Lee e Hahn (1994) os valores dos limiares são calculados para variar em cada tipo de locução. Para esse cálculo, naquele estudo foi pressuposto que o ambiente acústico não muda significantemente em um curto intervalo de tempo (aproximadamente 1.3 s), o maior tempo necessário para a locução de uma palavra no vocabulário coreano. Cada sinal de entrada de fala era divido em quadros de 10 ms de comprimento. O nível DC era primeiramente calculado para os três quadros iniciais e era usado para eliminar o viés do DC dos quadros seguintes. Para os três quadros seguintes (4-6), o valor do limiar para a energia era estimado, bem como os valores de *Max-plus e Max-minus*, que são usados para o cálculo do LCR fixados. Usando estes *Max-plus e Max-minus*, o limiar para o LCR dos três quadros seguintes (7-9) foi avaliado. Isto significa que eram necessários pelo menos 100 ms de intervalo de silêncio antes que cada locução real fosse presumida. O valor do limiar deste intervalo era primeiramente calculado e, então, a categorização de cada quadro seguinte como silêncio ou fala era executada. As características descritas acima foram calculadas da seguinte forma:

Energia = LCR aumenta em 1 quando:
(S(n-1) > Max_plus && S(n) < Max_minus||
(S(n-1) < Max_minus && S(n) > Max_plus).

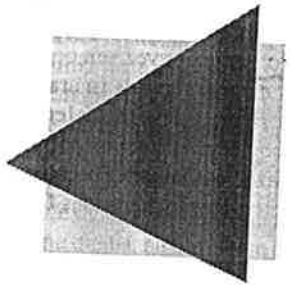
A decisão quanto ao início e fim da locução era baseada na seguinte regra: Quando 5 quadros consecutivos eram classificados como fala, o primeiro quadro do intervalo era definido como o começo, e quando mais de 13 quadros eram classificados como silêncio em 15 quadros consecutivos, o último era marcado como o fim.

Para uso eficiente de memória, eram armazenadas apenas parte das falas pré-determinadas de todas as locuções que davam entrada na memória. E a classificação fala/silêncio de cada quadro era delineada para ser feita antes de cada interrupção para a conversão A/D (analógico/digital) a fim de garantir a propriedade de tempo real. Em outras palavras, se o início do endereço de memória fosse 30001, 300 dados, isto é, 3 quadros, eram coletados e armazenados acima do endereço 30300, e o valor de *dc_offset* era computado para este dado durante a *interrupt*, e então o ponteiro do endereço era *re-setado* para 30001. Da mesma forma, após ter configurado os valores dos limiares de energia e LCR, os dados de entrada de um quadro de comprimento, i.e., 100 pontos de dados, eram armazenados no *buffer* de memória pré-reservado. Finalmente, a decisão fala/silêncio era tomada durante o intervalo de interrupção. Se o quadro fosse classificado como fala, o dado era armazenado na memória, caso contrário, era descartado. Assim, se houvesse apenas 2 s de fala em 20 s de dados, eram necessários apenas 40 Kbyte de espaço de memória para armazenar o dado de fala.

O PC controlava todas as operações do algoritmo (ele inicializava a placa de som, decifrava os códigos, executava os programas, e carregava os dados quando a execução do programa era finalizada). A classificação fala/silêncio era feita totalmente na placa de som. Em outras palavras, o algoritmo aceitava a entrada de dados de fala através de um canal interno A/D usando uma técnica de interrupção do *timer*, e ao mesmo tempo, fazia a classificação fala/silêncio e armazenava apenas parte dos dados classificados como fala, e então informava a execução de um programa para o PC.

A avaliação de desempenho era feita a partir da comparação dos resultados dos algoritmos com os da segmentação manual. A segmentação manual era executada a partir da inspeção cuidadosa do sinal original e do pré-determinado. Os resultados de detecção de início e fim de locução obtidos por meio do algoritmo de Lee e Hahn, em comparação com os de Rabiner e Sambur, e de Hahn, deixa claro que o primeiro é o melhor. As médias de erros foram 4.9 ms para o de Lee e Hahn, 7.6 ms para o de Hahn, e 10.9 ms para o de Rabiner e Sambur. Assim o algoritmo de Lee e Hahn foi superior aos outros dois outros na detecção do início das fricativas, e trabalhou muito melhor em intervalos com ausência de voz do que os outros dois. Muitos dos erros de detecção de final de locução em seu algoritmo derivavam de erros que ocorriam nos intervalos de silêncio.

Portanto, o algoritmo de Lee e Hahn é razoavelmente simples e desempenha com sucesso a detecção de fala em locução de palavras isoladas. A avaliação do desempenho dos participantes mostra que o algoritmo trabalha muito bem e é muito superior aos de Rabiner e Sambur, e ao de Hahn. Isto permite seu uso em situações reais, especialmente para sistemas baseados em DTW de IWR em tempo real, e em sistemas automáticos de coleta de dados de fala. Os dados mostram também que seu erro médio é menor que o comprimento de um quadro. Em conseqüência, é eficaz para garantir um bem sucedido alinhamento de tempo entre uma entrada desconhecida e um modelo pré-gravado quando um DTW é adotado. Os autores admitem que seu algoritmo tem que ser testado com uma base maior de dados incluindo locuções femininas a fim de confirmar sua utilidade em várias situações, e crêem que o uso de algumas características de tempo real para os sinais de fala pré-determinados melhorará seu desempenho.



ANEXO 2

ALGUMAS PERSPECTIVAS DE APLICAÇÃO DE CRONOFONOS 2.0 PARA AVALIAÇÃO DE DESENVOLVIMENTO DE LEITURA BEM COMO DE ANÁLISE DE PADRÕES DE DÉFICITS DE LEITURA EM DISLEXIAS.

A proposta de emprego do *software* CRONOFONOS 2.0 para avaliação de desenvolvimento de leitura e de análise de padrões de déficit de leitura em dislexias foi descrita brevemente em Capovilla e cols. (1996). As dislexias são distúrbios específicos de leitura que podem ser adquiridos ou do desenvolvimento. Nas dislexias adquiridas há uma perda da leitura em conseqüência de lesão cerebral, que pode acompanhar quadros de perda de linguagem falada (afasias). Nas dislexias de desenvolvimento há dificuldades na aquisição de leitura, que são maiores do que aquelas compatíveis com a idade e inteligência da criança. A dislexia do desenvolvimento pode ser vista como uma interrupção no desenvolvimento normal das habilidades de leitura e escrita (Morton, 1989). Tal desenvolvimento ocorre ao longo dos estágios logográfico, alfabético, e ortográfico (Frith, 1985). A interrupção afeta o desenvolvimento das habilidades alfabéticas e ortográficas. Quando as habilidades de leitura de disléxicos são comparadas àquelas de crianças normais, ficam evidentes disfunções básicas no sistema fonológico, que se manifestam em reduzida habilidade de manipulação fonológica. Assim, tais crianças tendem a ter dificuldade em reconhecer e produzir rimas e aliteração; em reconhecer e corrigir erros de pronúncia; em contar segmentos de palavras ouvidas; em dizer como soariam palavras sem uma parte delas ("cabelo" sem "be"), ou com uma outra parte adicionada no início, meio, ou fim ("carrão" com "ma" no início), etc. Tem sido demonstrado que a habilidade de manipular sons está claramente relacionada às habilidades de leitura e ditado em línguas alfabéticas como inglês (Warrick, Rubin, & Walsh, 1993) e português (Capovilla e cols., 1995).

Dentre todos os modelos teóricos para compreensão da dislexia, o mais bem sucedido é o Modelo de Duplo Processo (Marshall, 1989), baseado na Teoria de Processamento de Informações. O *software* por nós desenvolvido, baseado naquele modelo, nos dados de Pinheiro (1994), e de Capovilla e cols. (1995). O *software* CRONOFONOS 2.0 objetiva diagnosticar o tipo de dislexia apresentado pela criança, para permitir intervir de maneira concentrada e focal. No Brasil há uma carência de instrumentos de diagnóstico e tratamento, especialmente instrumentos computadorizados baseados em dados normativos. CRONOFONOS 2.0 permite analisar o grau de desenvolvimento de leitura de palavras isoladas em alfabetizandos, bem como diagnosticar diferencialmente o tipo de dislexia de desenvolvimento que uma criança

apresenta. Ele consiste em um exame computadorizado que é apresentado à criança sob a forma de um jogo. Neste "jogo" ocorre o seguinte: Dezenas de fotos coloridas e palavras escritas em tamanho grande aparecem ao monitor, uma por vez, em ordem aleatorizada. Quando surge uma foto, a criança deve tentar dizer o nome da figura em voz alta. Do mesmo modo, quando aparece uma palavra escrita, ela deve tentar ler em voz alta o que está escrito. A criança usa um microfone acoplado ao computador por uma placa de som. O *software* permite ao computador fazer a comparação entre a verbalização da criança diante de palavras escritas e sua verbalização diante das figuras que as representam.

Crianças com distúrbios específicos de leitura conseguem nomear as figuras mas não conseguem ler as palavras correspondentes quando estas têm determinadas características, e acabam "saltando" as palavras escritas com que têm dificuldade ou vocalizando-as incorretamente na leitura em voz alta. O computador sempre compara os sons das palavras faladas pela criança durante a nomeação com aqueles das palavras faladas por ela durante a leitura. Assim, o computador "sabe" se a palavra falada pela criança durante a leitura em voz alta corresponde ou não à palavra correta (isto é, aquela que a criança falou durante a nomeação). Além disso, quando a criança tem dificuldade com palavras com determinadas características, ela tende a apresentar hesitação diante dessas palavras escritas. Quando isto ocorre, o computador também "percebe" a dificuldade, já que ele compara a latência, a duração, e os intervalos da vocalização frente a palavras diferentes com aqueles frente às figuras correspondentes.

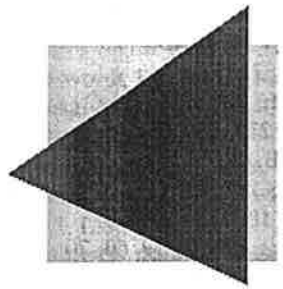
CronoFonos 2.0 analisa os padrões de leitura da criança em termos tanto de erros como de tempo despendido, com precisão de ms. Tanto nas crianças normais quanto nas disléxicas, os erros cometidos na leitura e o tempo despendido na leitura variam dependendo de certas características das palavras a serem lidas. O *software* apresenta à criança palavras escritas com diferentes características psicolinguísticas de modo a definir o grau de desenvolvimento de uma ou outra rota de leitura, bem como o tipo de dislexia que a criança apresenta. De modo geral há basicamente duas rotas de leitura: fonológica e lexical. Na leitura fonológica a criança lê por decodificação seqüencial das partes, relacionando os segmentos escritos aos seus respectivos sons. Na leitura lexical a criança lê a palavra por inteiro, de uma vez. A rota fonológica é mais lenta: a latência é maior, a duração é maior, e os intervalos entre as sílabas são mais longos. Por meio dela a criança pode ler palavras reais, palavras novas, ou mesmo inventadas, desde que sejam regulares em termos de relação grafema-fonema. A rota lexical é mais rápida. Por meio dela a criança pode ler palavras regulares ou irregulares, desde que sejam reais e a criança tenha bastante familiaridade com elas (isto é, desde que elas tenham relativamente alta frequência de ocorrência na língua). O registro das características específicas das palavras que representam dificuldade para a criança permite determinar as rotas empregadas. São quatro as características relevantes: lexicalidade, regularidade, frequência de ocorrência, e comprimento. Isto é melhor descrito abaixo.

Os itens que a criança deve ler variam em termos de sua lexicalidade. Alguns são palavras reais como *pato*; outros são itens "inventados", também chamados de pseudo-palavras, e não têm sentido como *pota*. As pseudo-palavras não podem ser lidas lexicalmente (isto é, por inteiro, de uma vez), mas apenas por decodificação fonológica (isto é, por partes, relacionando segmentos escritos com seus

sons correspondentes). As palavras variam também em sua regularidade: algumas têm relações grafema-fonema exclusivamente regulares, ou seja, para cada grafema há um fonema e vice-versa. Um exemplo é *pata*. Outras palavras envolvem alguma *regra de posição* , como por exemplo a regra: "o *s* intervocálico soa como */z/*") como na palavra *casa*. Outras palavras ainda contêm relações grafema-fonema irregulares, ou seja, para um mesmo fonema há vários grafemas e vice-versa. Um exemplo é a palavra *táxi*. As palavras irregulares não podem ser lidas fonologicamente, mas apenas lexicalmente. Finalmente as palavras variam também em frequência de ocorrência na língua: algumas têm alta frequência como *gato*; outras, baixa como *unha*. Quanto maior a frequência de ocorrência, tanto maior a probabilidade de uma palavra ser lida lexicalmente.

Embora as dificuldades apresentadas por disléxicos sejam consideravelmente mais severas do que aquelas apresentadas por crianças normais, ainda assim a natureza da dificuldade está relacionada às características psicolinguísticas em casos bem definidos de dislexia. Características de palavras que obstaculizam a leitura em certas dislexias são inócuas em outras. Assim, crianças cuja dislexia advém de distúrbios fonológicos (isto é, perda da habilidade de converter segmentos escritos em segmentos falados) não lêem fonologicamente mas apenas lexicalmente; portanto elas têm dificuldades em ler pseudo-palavras, ou palavras reais de baixa frequência. Já crianças que têm distúrbios lexicais não lêem lexicalmente, mas apenas fonologicamente; portanto, elas têm dificuldades com palavras irregulares, mas são menos afetadas pela presença ou não de significado na palavra.

CronoFonos 2.0 permite analisar o amadurecimento das rotas de leitura em crianças com desenvolvimento normal, além de servir de importante auxílio na identificação precisa do tipo e do grau de dislexia apresentados. Isto é o primeiro e mais importante passo para assegurar uma efetiva intervenção focalizada nas dificuldades específicas apresentadas pela criança. Ele é também um instrumento sensível e preciso para avaliar o efeito de diferentes programas de intervenção em remediação de distúrbios de leitura. Isto representa uma contribuição de inestimável valor para elevar o rigor metodológico e o teor de cientificidade das pesquisas no campo. CronoFonos 2.0 representa uma ferramenta para pesquisa, avaliação e tratamento na área de desenvolvimento de leitura e de seus distúrbios.



Índice por Autor

Aday, L.A., 233, 256
Algarabel, S., 195
Algina, J., 18, 22, 200, 207, 208
Aligieri, S., 281
Allegiant, 214, 227
Almeida, L.S., 206
Altose, M.D., 91, 100
Alves Pereira, C.A., 131, 135, 140
Alwin, D.F., 121, 124
American Psychiatric Association, 30, 94, 98
Anastasi, A., 18, 21, 27, 34, 59, 70, 180, 202, 206
Andrade, M.P., 280
Anes, M.D., 282
Angoff, W.H., 18, 21
APA, 17, 21, 209, 214, 215, 227, 239, 244, 256
Applied Psychological Measurement, 18, 21
Araújo, I., 281
Arter, J.A., 18, 23
Ashby, F.G., 17, 25
Assessment Systems Corporation, 214, 223, 224, 227
Ato, M., 212, 228
Austria, A.M., 58, 71, 127
Babbie, E., 233, 250
Backhoff, E., 210, 227
Bacon, F., 14, 21
Baird, J.C., 74, 81, 91, 98, 100
Ball, E.W., 278
Ball, L.C., 284
Ballin, M., 118, 124
Barbero, M., 221, 223, 227
Barclay, J.R., 14, 21
Barnes, L.B., 213, 230
Bartram, D., 213, 227
Baxter, G.P., 145, 182
Beck, I., 284
Behlau, M.S., 259, 268, 277, 278, 279, 285
Bejar, I.I., 226, 227
Bell, S.R., 18, 25
Bendig, A.W., 58, 70, 125
Berk, R.A., 18, 21
Bernat, A.B., 281
Berry, D., 283
Binet, A., 15, 16, 20, 21, 22, 180
Bingham, W.V.D., 236, 256
Bird, C., 119, 125
Birnbaum, A., 64, 70
Birnbaum, M.C., 86, 100
Birnbaum, M.H., 88, 98
Blachman, B.A., 278
Black, H.D., 204, 205, 206
Block, J.H., 194, 196, 203, 207
Bloom, B.S., 145, 149, 163, 164, 179, 180, 181, 193, 207
Bock, R.D., 223, 229
Bohannon, M.W., 87, 88, 103
Bolton, T.L., 20, 22
Booth, P., 214, 227
Borg, G., 92, 99

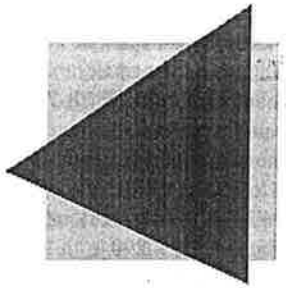
Borich, G., 181
Boring, E.G., 14, 16, 22
Bortz, J., 251, 252, 256
Bowie, D.M., 92, 101
Bradburn, N.M., 242, 244, 258
Brady, S., 284, 285
Brand, N., 213, 227
Braun, H.I., 18, 25
Brennan, E.M., 89, 99
Brinker, R.P., 83, 86, 89, 99
Brito, S.M.O., 235, 237, 256
Brooks, R.M., 255, 256
Brown, F.G., 18, 22
Brown, R.T., 18, 23
Bunderson, C.V., 210, 211, 216, 221, 227
Burdon, J.G.W., 92, 99
Buros, O.K., 17, 22
Burt, C., 17, 22
Byrd, D., 268, 279
Byrne, B., 279
Campbell, D.T., 28, 34, 62, 67, 70
Campbell, E.J.M., 99, 101
Capovilla, A.G.S., 263, 279, 280, 281, 282, 283
Capovilla, F.C., 261, 263, 265, 269, 279, 280, 281, 282, 283, 284, 286, 291
Carmines, E.G., 111, 121, 125, 126, 127
Carpenter, E.H., 255, 256
Carpenter, P.A., 18, 22
Carroll, J.B., 211, 227
Carvalho, V., 281
Carver, R.P., 191, 207
Cattell, J.M., 16, 20, 22
Cattell, R.B., 17, 22, 27, 34
César, O.P., 263, 279
Chave, E.J., 17, 25, 117, 127
Christenson, J.A., 255, 256
Clausen, A.R., 111, 124
Cluff, M., 282
Coffman, W.E., 175, 181
Cohen, J., 200, 207
Cole, N.S., 223, 227
Colorni, E.M.R., 263, 279, 280
Conoley, J.C., 211, 229
Coombs, C.H., 124, 125
Corga, D., 257
Correia, L.M., 206
Costa, C.E., 279, 280, 281
Coulson, D.B., 207
Crandall, C.S., 88, 99
Crocker, L., 18, 22
Cronbach, L.J., 32, 34, 59, 69, 70
Crooks, T.J., 145, 180
Cross, D., 85, 101
Cross, D.V., 80, 99
Cunha, W.H.A., 33, 34
Da Silva, J.A., 73, 75, 76, 91, 99, 100
Daubenspeck, J.A., 91, 101
Davies, I.K., 181
Dawson, W.E., 83, 86, 89, 99
Delbeke, L., 124, 125
Delgado, A.P., 261, 283
Delgado, A.R., 215, 229
Detterman, D.K., 18, 24
Devine, D., 213, 230
Diamico, K., 257
Dillman, D.A., 236, 255, 256, 257
Dockrell, W.B., 205, 206
Döring, N., 251, 252, 256
Drasgow, F., 60, 70, 181
Dubois, P.H., 14, 19, 22
Duduchi, M., 261, 263, 265, 269, 279, 280, 281, 282, 283, 284
Dunn, L.M., 282
Dunn, T., 209, 228
Ebbinghaus, H., 20, 22
Ebel, R.L., 202, 207
Edwards, A.L., 110, 112, 117, 118, 121, 125
Ehri, L.C., 283
Eignor, 189
Ekman, G., 81, 82, 83, 84, 99, 124, 125
Ellis, A., 260, 282
Ellis, L.W., 89, 99, 100
Embretson, S., 226, 227
Engelman, A., 75, 80, 99
Esquirol, J.E.D., 19, 22
Evan, W.M., 213, 227
Eysenck, M.W., 260, 282
Faleiros Sousa, F.A.E., 73, 86, 89, 100
Farnsworth, P.R., 118, 124
Feitosa, M.D., 263, 281, 282, 284, 286
Feltham, R., 213, 229
Ferraz, A., 257
Ferreira, F., 282
Fidell, L.S., 252, 258
Fielding-Barnsley, R., 279
Fink, A., 34, 231, 243, 256
Fischer, G.H., 226, 228
Fiske, D.W., 62, 70
Fitz-Gibbon, 150, 151
Flood, M.A., 87, 101
Foley, H., 101
Foley, M.A., 101
Fowler, F.J., 242, 256
Fowler, R.D., 212, 228
Françozo, E., 260, 276, 282
Frankel, M., 233, 256
Fraser-Robinson, J., 255, 256
Frederiksen, J.R., 226, 228
French, W., 163, 181
Frey, J.H., 255, 256
Frith, U., 260, 282, 291
Fucci, D.J., 89, 99, 100
Fukusima, S.S., 75, 91, 99, 100
Fuso, S.F., 281
Gagné, R.M., 145, 181
Gaito, J., 17, 22
Galaburda, A.M., 282
Galton, F., 15, 16, 19, 22
Ganança, M.M., 268, 278, 279
Geller, D., 90, 100
Gescheider, G.A., 74, 75, 100
Gielow, I., 282
Gilbert, J.A., 20, 22
Gilliland, S.W., 213, 230
Glaser, R., 32, 35, 189, 192, 202, 203, 204, 205, 207, 226, 229
Glover, J.A., 211, 229
Goldsamt, M.R., 58, 70, 125
Goldshalk, F.I., 175, 181
Goleman, D., 146, 181
Gonçalves, M.J., 261, 263, 280, 281, 282, 283, 284
Gottfried, S.B., 91, 100
Gough, P.B., 283
Gouveia, V.V., 238, 256
Gray, W.M., 189, 207
Green, B.F., 215, 228
Green, S.B., 145, 181
Greenbaum, H.B., 102
Greene, V.L., 121, 125
Greeno, J.G., 145, 181
Gress, L.D., 87, 101
Griffith, P.L., 283
Griswold, M.M., 145, 182
Gronlund, N.E., 18, 22, 157, 181, 190, 191, 193, 197, 201, 203, 207
Guedes, M., 280, 281
Guilford, J.P., 17, 22, 110, 125
Guilherme, R.L., 283
Guirao, M., 77, 102
Gulliksen, H., 17, 22, 59, 70
Günther, H., 235, 237, 238, 243, 256
Guttman, L., 118, 125
Hahn, M., 265, 266, 283, 287, 288
Haladyna, T.M18, 23, 145, 153, 154, 181, 182
Halpin, G., 145, 181
Halpin, G.W., 145, 181
Hambleton, R.K., 18, 23, 59, 63, 70, 189, 192, 194, 199, 200, 201, 202, 203, 207, 208, 214, 215, 218, 220, 223, 224, 226, 228, 230
Hamblin, R.H., 86, 100
Hamblin, R.L., 86, 100
Hansen, D.N., 213, 228
Harari, H., 232, 258
Hardin, C., 86, 100
Hargreave, F.E., 92, 99
Harman, H.H., 17, 23, 60, 70
Harrison, R.C., 232, 258
Harrow, A.J., 148, 181
Harver, A., 91, 100, 101
Harvey, A.L., 212, 213, 228, 230
Hastings, J.T., 145, 180
Haydu, V.B., 263, 280, 283
Held, J.J., 213, 228
Henderson, J., 282
Henerson, 179, 181

Henri, V., 20, 21
Henry, N.B., 159, 181
Henry, G.T., 233, 257
Herzel, H., 283
Hevner, K., 118, 126
Hinshaw, 98
Hohn, W.E., 283
Holland, P.W., 18, 23
Holmes, T.H., 87, 88, 100, 101, 102, 103
Holst, P.M., 225, 229
Hosogi, M.L., 283
Houx, P.J., 213, 227
Howard, K.I., 58, 127
Hughes, C., 284
Hulin, C.L., 60, 70, 181
Hume, D., 14, 23
Husek, T.R., 192, 207
Ibarra, M.A., 210, 227
Indow, T., 82, 83, 100
Inouye, D.K., 210, 211, 216, 227
Ippel, M.J., 211, 217, 228
Jablonski, B., 257
Jacob, A.N., 283
Jacoby, J., 58, 71, 126, 251, 257
Jenkins, J.J., 126, 140
Jenkins, J.R., 284
Johnson, 225
Joncich, G., 15, 23
Jones, N.L., 92, 101
Jones, R.R., 58, 70, 126
Judd, Ch.M., 233, 257
Juel, C., 283
Juniper, E.F., 92, 99
Just, M.A., 18, 22
Kail, R.V., 226, 229
Kalton, G., 232, 257
Kamisaki, R., 88
Karasu, T.B., 213, 229
Keane, M.T., 260, 262
Kelley, T.L., 17, 23, 181
Kemp, S., 84, 101
Kenny, K.C., 121, 125
Keyser, D.J., 217, 230
Kidder, L.H., 233, 257
Kilkenny, M.J., 244, 258
Killian, K.J., 92, 99, 101
Kilpatrick, F.P., 118, 125
Kim, M.J., 91, 92, 101
Kish, L., 231, 233, 257
Kjaer, G., 87, 102
Kline, P., 27, 35
Knight, K.K., 90, 101
Kohn, S.D., 82, 101
Komorita, S.S., 58, 71, 126
Kosecoff, J., 34, 231, 243, 256
Kotses, H., 91, 100
Kraepelin, E., 20, 23, 29, 35
Krathwohl, D.R., 145, 181
Krosnick, J.A., 253, 255, 257
Krueger, L.E., 75, 101
Kruskal, J.B., 124, 126
Kubiszyn, T., 181
Kuder, G.F., 69, 71
Künnapas, T.M., 81, 82, 84, 86, 99, 101, 125
Kyllonen, P.C., 210, 211, 215, 228
Lamel, L.F., 283, 288
Landis, R.S., 213, 230
Larar, J.N., 283, 288
Laurendeau, M., 182
Lavrakas, P.J., 255, 257
Le Blanc, P., 92, 101
Lecours, A.R., 261, 283
Lee, H.S., 265, 266, 283, 287, 288
Leicester, N., 284
Lemann, I.J., 18, 23
Lemle, M., 259, 262, 283
Lentine, T., 91, 101
Levandowski, 268
Lewis, B., 263, 286
Lewis, C., 69, 71
Likert, R., 50, 71, 119, 120, 121, 126
Lincheid, T.R., 94, 103
Linn, R.L., 189
Lissitz, R.W., 18, 24
Lobel, S.A., 257
Lodge, M., 80, 81, 85, 101
Lohman, D.F., 145, 182, 211, 217, 226, 228, 230
Lopes, Jr., J., 243, 256
Lopez, J.A., 212, 228

Lord, F.M., 17, 23, 60, 64, 71, 181, 222, 223, 228, 230
Luce, P., 282
Luce, R.D., 109, 110, 112, 126
Lugo, D.E., 282
Lundberg, I., 284
Lushene, R.E., 209, 228
Luzia, J.C., 280
Lynn, E., 232, 258
Macedo, E.C., 261, 263, 265, 269, 279, 280, 281, 282, 283, 284, 286
Macedo, L., 75, 76, 99
Mack, J.D., 78, 102
Madaus, G.F., 145, 180
Mager, R.F., 46, 48, 71, 144, 180, 182, 183, 188
Maher, M.P., 240, 258
Mahler, D.A., 91, 101
Mahutte, C.K., 101
Malle, B.F., 211, 217, 229
Mangione, Th.W., 239, 257
Mann, V., 284
Margolis, R.H., 90, 100, 101
Marshall, J., 284, 291
Masia, B.B., 145, 181
Masuda, M., 87, 88, 101, 103
Matell, M.S., 58, 71, 126, 251, 257
Mayer, D., 112
Mayo, E., 32, 35
Mazzeo, J., 212, 228
McClean, 268
McClelland, J.L., 260, 285
McFarlane, D.K., 282
McGovern, J.F., 91, 101
McIver, J.P., 111, 121, 126
Mednick, S.A., 28, 35
Meehl, P.E., 32, 34
Mehrens, W.A., 18, 23
Messick, S., 124, 127, 145, 182, 207
Metz, D.E., 91, 102
Metz-Lutz, M.N., 260, 284
Meyer, M., 102, 126
Michell, J., 17, 23
Milgram, S., 249, 257
Milenkovic, P., 263, 268, 284
Miller, D.G., 285
Miller, J.R., 213, 227
Millmán, J., 18, 23, 202, 207
Mirando, M.A., 86, 89, 99,
Mislevy, R.J., 223, 229
Monnerat, M., 257
Moore, B.V., 236, 256
Moore, C.F., 98, 103
Moos, R.H., 251, 257
Moreira, M.A.C., 281
Morris, C.W., 129, 140, 150, 151, 179
Morton, J., 259, 260, 261, 284, 291
Moss, P.A., 227
Muheenkamp, A.F., 87, 101
Mumaw, R.J., 226, 229
Muñiz, J., 60, 63, 71, 218, 220, 222, 229
Murdaugh, 98
Neubauer, A.C., 211, 217, 229
Neuburg, E.P., 284, 288
Nevo, B., 50, 71
Nico, A.M., 280
Nield, M., 91, 92, 101
Nitko, A.J., 189, 199, 202, 207
Nogueira, D., 281
Noma, E., 74, 81, 98
Novick, M.R., 17, 23, 69, 71, 192, 199, 207
Nunes, D., 263, 281
Nunes, L.R.O.P., 263
Nunnally, J.C., Jr., 59, 71, 121, 126
O'Connor, R.E., 284
O'Donnell, A.M., 225, 229
Olea, J., 224, 229
Olsen, J.B., 210, 211, 216, 227
O'Neil, H.F., 213, 228
O'Neil, O.F., 209, 228
Osgood, C.E., 129, 130, 131, 133, 135, 140
Osterling, S.J., 18, 23
Ottonson, D., 92, 99
Padilla, E.R., 282
Pareek, U., 235, 257
Parente, M.A.P., 283, 284
Park, C.K., 283, 288
Parsons, C.K., 60, 70, 181
Pasquali, L., 60, 62, 71, 110, 111, 220,

229, 247, 257
Patel, M., 91, 92, 101
Pearson, K., 15
Pellegrino, J.W., 226, 229
Pereira, M., 257
Perfetti, C.A., 284
Perloe, S.I., 86, 102
Petrosino, L., 89, 100
Petz, B., 112, 126
Piaget, A., 182
Pieters, J.P.M., 214, 215, 228
Pimenta, M.A.M., 261, 283
Pimentel, N.S., 283
Pinard, A., 182
Pine, J., 145, 182
Pinheiro, A.M.V., 259, 260, 261, 264, 268, 276, 278, 285, 291
Plake, B.S., 213, 225, 230
Plutchik, R., 213, 229
Ponsoda, V., 224, 229
Pontes, P.A., 268, 278, 279
Popham, W.J., 189, 192, 207, 208
Popper, K.R., 14, 23
Poulton, E.C., 75, 102
Powell, 189
Pratt, A., 285
Presser, S., 243, 257
Prieto, G., 215, 229
Quetelet, 15
Rabiner, L.R., 265, 285, 286, 287, 288
Rahe, R.H., 87, 100, 102
Rao, T.V., 235, 257
Raphael, W.D., 263, 280, 281
Rasch, G., 63, 71
Rechtschaffen, A., 28, 35
Reckase, M.D., 226, 229
Redline, S., 91, 100
Reeder, R., 85, 101
Renom, J., 224, 225, 229
Reynolds, C.R., 18, 23
Rhodes, I.N., 244, 258
Ribeiro, A., 281
Ribeiro, G., 75, 100
Ribeiro, I.S., 206
Richardson, M.W., 69, 71
Rifkin, B., 24
Rocklin, T.R., 225, 229
Rodeghier, M., 238, 257
Rodrigues, A., 255, 257
Rogers, H.J., 218, 220, 223, 228
Rogers, J., 18, 23, 59, 63, 70
Rohl, M., 285
Roid, G.H., 18, 23, 145, 182
Rolls, S., 213, 229
Ronning, R.R., 211, 229
Rosas, M., 210, 227
Rosenthal, R., 263, 285
Rosiello, R.A., 91, 101
Rosnow, R.L., 263, 285
Ross, 225
Rowe-Walsh, S., 286
Rubin, D.B., 18, 23
Rubin, H., 286, 291
Ruch, L.O., 87, 102
Rulon, P.J., 69, 71
Russell, W.A., 125, 140
Russo, I., 259, 268, 277, 285
Ryan, E.B., 89, 99
Saffir, M.A., 118, 126
Saleh, M., 283
Sambur, M.R., 265, 285, 288
Sanchez, J., 212, 228
Santisteban, C., 60, 71
Santos, A., 281
Santos, M.T.M., 285
Schafer, L.W., 285, 287
Schaffer, D.R., 255, 257
Schepp, K.G., 98, 102
Schiavetti, N., 91, 102
Schmitt, N., 213, 230
Schoonman, W., 213, 225, 230
Schuman, H., 232, 243, 257
Schutte, H.K., 285
Schwarz, R.M., 244, 258
Scientific Software, 124, 126
Seabra, A.G., 279, 281, 282, 283, 284, 286
Seashore, R.H., 118, 126
Seidenberg, M.S., 260, 285
Seliger, E., 226, 227
Sellin, J.T., 85, 102
Sennott-Miller, 98
Serón, X., 260, 285
Sewell, W.H., 121, 126
Seymour, P.H., 261, 285
Sfez, J., 50, 71
Shallice, T., 260, 285
Sharp, S.E., 20, 24
Shavelson, R.J., 145, 182
Shaycoft, M.F., 194, 208
Shell, P., 18, 22
Sherriffs, A.C., 28, 35
Shinn, Jr., A.M., 84, 102
Siegel, S., 247, 258
Silva, A.V. da., 247, 258
Silva, L.S., 280
Silva, M.S.M.M., 235, 237, 256
Simon, Th., 15, 16, 20, 22, 180
Singer, E., 240, 258
Sitler, R.W., 91, 102
Sjöberg, L., 125
Skaggs, G., 18, 24
Skurdal, M.A., 93, 103
Slocum, T.A., 284
Smith, C.R., 86, 100
Smith, E.R., 232, 257
Smith, M., 87, 102
Snow, R.E., 145, 182, 211, 226, 230
Solórzano, I.M., 238, 258
Sommer, B., 243, 247, 258
Sommer, R., 243, 247, 250, 258
Soria, R., 261, 263, 265, 269, 282, 284, 285
Sotoodeh, Y., 88, 98
Spearman, C., 15, 16, 24, 182
SPSS, Inc., 222, 230
Stanley, J.C., 67, 70
Stanovich, K.E., 285
Stenberger, J.G., 285
Sternberg, R.J18, 24, 182, 226, 230
Stevens, J.C., 78, 102
Stevens, S.S., 17, 24, 74, 75, 76, 77, 78, 79, 80, 81, 83, 90, 97, 102, 110, 126, 252, 258
Stiggins, R.J., 145, 182
Stone, L.A., 93, 94, 102, 103
Stone, C.W., 17, 24
Suci, G.J., 125, 129, 131, 133, 135, 140
Sudman, S., 233, 242, 244, 258
Sullivan, R., 92, 93. 103
Sulter, A.M., 285
Summers, E., 92, 101
Sundberg, N.D., 13. 24
Swaminathan, H., 18, 23, 59, 63, 70, 200, 207, 208, 218, 220, 223, 228
Sweetland, R.C., 217, 230
Swineford, F., 175, 181
Tabachnick, B.G., 252, 258
Tanenhaus, J., 85, 101
Tannenbaum, P.H., 129, 131, 133, 135, 140
Telage, K.M., 286
Telebrás, 184, 186, 187
Tenney, S.M., 91, 100
Terman, L.M., 16, 20, 24
Thalmann, R., 90, 103
Thiers, V.O., 263, 280, 281, 282, 283, 284, 286
Thomson, G.H., 17, 25
Thorndike, E.L., 15, 17, 25
Thorndike, R.L., 18, 25, 60, 71
Throneburg, R.N., 263, 268, 286
Thurstone, L.L., 15, 16, 17, 21, 25, 111, 112, 117, 126, 127
Thurstone, T.G., 25
Titze, I.R., 283
Torgerson, W.S., 17, 25, 124, 127, 220
Tosi, O., 268, 278, 279
Townsend, J.T., 17, 25
Treiman, R., 285
Tryon, W.W., 92, 103
Tucker, L.R., 124, 127
Tunmer, W.E., 285
Tursky, B., 85, 101
Tyler, R.W., 148, 163, 182
Van der Linden, W.J., 218, 220, 226, 230
Van der Veer, F., 58, 71, 127
Van Hoewyk, J., 240, 258
Van Horn, C.E., 111, 125
Velandrino, A.P., 212, 228
Viana, A.L.A.G., 281
Volicer, B.J., 87, 88, 103

Wainer, H., 18, 25, 221
Walsh, 291
Warburton, F.W., 17, 22, 27, 34
Warren, R.M., 75, 103
Warren, R.P., 75, 103
Warrick, N., 286, 291
Webb, S.C., 118, 127
Weeks, P.A., 282
Weil, E.M., 18, 24
Wiggins, G., 145, 182
Wikelund, K.R., 14, 182
Wikstroem, I., 86, 101
Williams, B., 233, 258
Wills, C.E., 98, 103
Wilpon, J.G., 286, 288
Wingersky, M.S., 222, 230
Wise, S.L., 213, 225, 230
Wish, M., 124, 126
Wissler, C., 20, 25
Wit, H.P., 285
Witt, J.C., 211, 229
Wolf, R.L., 279
Wolfgang, M.E., 85, 102
Wolpe, J., 92, 103
Wright, B.D., 18, 25
Wyler, A.R., 88, 103
Yairi, E., 263, 268, 286
Yaremko, R.K., 232, 258
Yela, M., 60, 71
Yen, W.M., 222, 230
Young, A.W., 260, 282
Zaal, J.N., 59, 70, 214, 215, 224, 228
Zdep, S.M., 244, 258
Zeller, R.A., 121, 125, 127



Índice por Assunto

Adequação do modelo (TRI), 222-223
Alfa de Cronbach, 69, veja Precisão
Amostra, 55-56, 196-197, 232-234
Análise
  Conceito, 169-170
  Técnicas de avaliação, 170-173
Análise de juízes, 53-54, veja Itens
Análise fatorial, 17, 60-62
Análise semântica, 52-53, veja Itens
Aplicação (objetivo educacional)
  Conceito, 163-164
  Técnicas de avaliação, 165-168
Aprendizagem, 141-143
Atributo, 39, 42-43
Atributos clínicos
  Ansiedade, 92
  Deficiências auditivas, 90-91
  Deficiências da fala, 89-90
  Diagnóstico psicopatológico, 93-94
  Dispnéia, 91-92
  Fobias, 92-93
  Gravidade de enfermidade, 88-89
  Medida, 87-94
  Stress, 87
Atributos sociais
  Bens públicos, 84-85
  Importância política, 81-82
  Julgamento moral, 83
  Liberalistmo vs. conservadorismo, 84
  Medida, 81-85
  Ofensas e crimes, 85
  Opniões sócio-políticas, 85
  Poder nacional, 84
  Preferência e prestígio profissionais, 86
  Preferências musicais, 82
  Preferências por relógio de pulso, 82-83
  Racismo, 83
  Status social, 86
  Valores estéticos, 82
Avaliação, 13, 141
Avaliação (objetivo educacional)
  Conceito, 175-176
  Formativa, 191
  Sumativa, 191-192
  Técnicas de avaliação, 176-179
Background (Survey)
  Cultural, 235
  Do pesquisador, 235
  Do respondente, 236
Banco de itens, 221-223
Bloom (taxionomia), 145-148
Complementaçao, veja Escala de resposta
Compreensão
  Conceito, 160
  Técnicas de avaliação, 160-162
Computador, veja Teste informatizado
Conceito, 232-234 (Survey), veja Significado (Diferencial semântico)

Conhecimento (objetivo educacional)
Conceito, 159
Técnicas de avaliação, 159-160
Consistência interna, 68-70, veja Precisão
Conteúdo, 141-142, 145
Contexto social (Survey), 235-236
Contínuo metatético e protético, 79-80
Correção Spearman-Brown, 68-69, veja Precisão
Critério, 202-204, veja CRT
Cronofonos
Dislexia, 291-293
Software, 265-268
Teste do, 268-278
CRT (criterion-referenced-test)
Características, 199-202
Conceito, 189-190
Construção, 192-198
Dificuldades, 202-204
Fidedignidade, 199-201
Normas, 201-202
Uso, 190-192
Validade, 201-202
Definição constitutiva, 44-45
Definição operacional, 45-47
Dislexia, veja Cronofonos
Diferencial semântico
Conceito, 131
Elaboração, 131-134
Diferença discriminante, 112
Dimensionalidade, 43-44, 60, 129-130
Distância semântica, 137-139
Conceito, 137
Fórmula, 138
Usos, 138-139
Domínio, 194-195, veja CRT
Domínio cognitivo, afetivo, psicomotor, Veja Bloom, Harrow
Domínio afetivo (técnicas de avaliação), 180
Duas metades, 68-69, veja Precisão
Dummy, veja Variável
Educação, 141-142
E-mail, veja Questionário
Emparelhamento, veja Escala de resposta
Emparelhamento intermodal, 78-79, veja Escala de razão
Emparelhamento numérico, veja Magnitude (estimação)
End-points (Cronofonos: determinação), 287-290
Ensaio, veja Escala de resposta
Ensino individualizado, 193
Entrevista
Individual, 254
Por telefone, 255
Escala
Bipolar, 131
Composição fatorial no DS, 135
Conceito, 105-106
De Fechner, 110
De Guttman, 108-110, 118-119
De intervalo, 107, 247, 249
De Likert, 108-110, 119-121, 250-252
De números, 106-107, 247-250, 252
De razão, 107, 247, 250, veja Escala de razão
De Stevens, 110, veja Lei da potência
De Thurstone, 108-110, 111-118
De Weber, 110
Formato no DS, 131
Multidimensional, 110-111, 121-124
Nominal, 107, 247-248
Ordinal, 107, 247-249
Psicofísica, 109-110, 112, veja Psicoffsica
Psicológica, 108-110, 111-114
Psicométrica, 109-110, 111-114
Unidimensional, 110-111
Escala de razão: elaboração
Aplicação da escala, 95-96
Emparelhamento intermodal, 96-97
Seleção de atributos, 94-95
Testes piloto, 95
Escala de resposta, 56-59
Complementação de sentença, 156-157
Emparelhamento, 156
Ensaio, 157-159
Escalas Likert, 56-59, 250-252
Escolha forçada, 56-59
Múltiplas alternativas, 57-58, 153-155
Portfólio, 159
Verdadeiro – falso, 155-156
Escalas psicofísicas, 34, veja Psicofísica
Escalas Likert, veja Escala e Escala de resposta,
Escolha forçada, veja Escala de resposta
Espaço semântico, 129-130, 137
Estimulação concorrente, 253
Estrutura conceitual, 139-140
Falso-negativo, 264
Fechner, veja Escala
Fidedignidade, veja Precisão
Fonema, 261-262
Formas alternativas, 68, veja Precisão
Gagné, 185-188
Grafema, 261-262
Guttman, veja Escala
Habilidades, 145
Harrow (domínio psicomotor), 148
Instruções, 58, 133-134
Internet, veja Questionário
Itens
Ameaçadores, 243-244
Análise empírica, 62-66, TRI: 221-223
Análise teórica, 52-54
Apresentação, 253-254
Chute, 63-65
Conceito, 47
De conhecimento, 244-245
De opinião, 245-246
Dificuldade ideal, 65-66
Dificuldade, 63-66
Discriminação, 63-65
Fatuais, 246
Fontes, 47-48
Grau de ameaça, 243-244
No survey, 233, 243-246
Quantidade, 51-52
Regras de construção, 48-51
Kuder-Richardson, 69, veja Precisão
Lei da potência, 73-76
Lei do julgamento comparativo, 113-114
Leitura, 259-265
Duração, 260-261
Fonológica e lexical, 259
Latência, 260-261
Rota fonológica, 261-264
Rota lexcical, 261-264
Segmento locucional, 261-264
Likert, veja Escala, Escala de resposta
Mager, 183-185
Magnitude, veja Psicofísica
Estimação, 76-77, 96-97
Produção, 77-78, 96-97
Mestria, 190-192, 202-204, veja CRT
Método dos intervalos aparentemente iguais, 117-118
Módulo livre, veja Magnitude (estimação)
Múltipla escolha, veja Escala de resposta
Notação manual vs. computadorizada, 263
Objetivos educacionais, 141, 143-148, 150-152, 193, 195, veja Bloom, Mager, Gagné
Observação de comportamento, 33, 231
Padrão de desempenho, 193-196
Pergunta aberta vs. fechada, 243
Pólo analítico, 38, 59-70
Pólo empírico, 37, 55-59
Pólo teórico, 37, 38-54
População, 232-234
Portfólio, veja Escala de resposta
Positivismo, 14-15
Precisão, 66-70
Procedimentos teóricos, experimentais, analíticos (veja

Pólo teórico, empírico, analítico)
Processo discriminante, 112
Processo mediativo, 129
Processos cognitivos, 144-145, veja Bloom, Harrow
Propriedade, veja Atributo
Pseudo-palavra, 259, 264
Psicofísica, cap. 4
e fenômenos subjetivos e sociais, 73, 80-81, veja atributos sociais e clínicos
Vantagens, 97-98
Psicometria, veja Escalas (escala psicométrica)
Questionário, 231
Administrador, 253-254
Aplicação via correio, 255
Aplicação via e-mail e internet, 255
Background conceitual, 232-234
Benefícios e recompensas, 240-241
Conceito, 231
Contexto social, 235-236
Custos, 238-240
Elementos, 242-252
Estrutura, 236-242
Pessoas envolvidas, 253-254
Princípios, 240-242
Rapport, 237-240
Respondente, 253-254
Tipos, 252-255
Rapport, veja Questionário
Representatividade dos itens, 202, 203
Reprodutibilidade, 119
Rota lexical, veja Leitura
Rota fonológica, veja Leitura
Segmento locucional, veja Leitura
Significado, 129, 261-262
Significado de um conceito, 136
Síntese (objetivo educacional)
Conceito, 173-174
Técnicas de avaliação, 174-175
Sistema psicológico, 40-42
Software, veja Cronofonos
Stevens, veja Lei da potência, Escala
Survey, 34, veja questionário
Tabela de especificação, 141, 148-152
Taxonomia (importância), 27
Temperamento, 13-14
Teoria de resposta ao item, veja TRI
Teoria psicológica, 38-39, veja Dimensionalidade
Teste adaptativo, 217, 223-224, 225
Teste centrado em critério, veja CRT
Teste informatizado
Apuração rápida, 212
Armazenamento rápido e grande, 211
Conceito, 209
Confiabilidade, 211
Economia, 210
Interação com examinando, 210, 212-215
Padronização, 210
Primeira geração (convencional), 216-217
Relato instantâneo, 212
Riqueza de estímulos, 211
Segurança, 210
Teste mental, 16, 20
Teste-reteste, 67-68, veja Precisão
Testes comportamentais, 33
Testes de inteligência, 16-17
Testes psicológicos, 19-21
Testes referentes a construto, 31-32, veja cap. 3,
Testes referentes a conteúdo, 32-33, veja cap. 7
Testes referentes a critério, 28-31
Thurstone, veja Escala
Treinamento, 141-142
TRI, 18, 63-65, 218-221, 225-226
Validade, 60-62
Valores, 179-180
Variável dummy, 252
Verdadeiro - falso, veja Escala de resposta
Weber, veja Escala