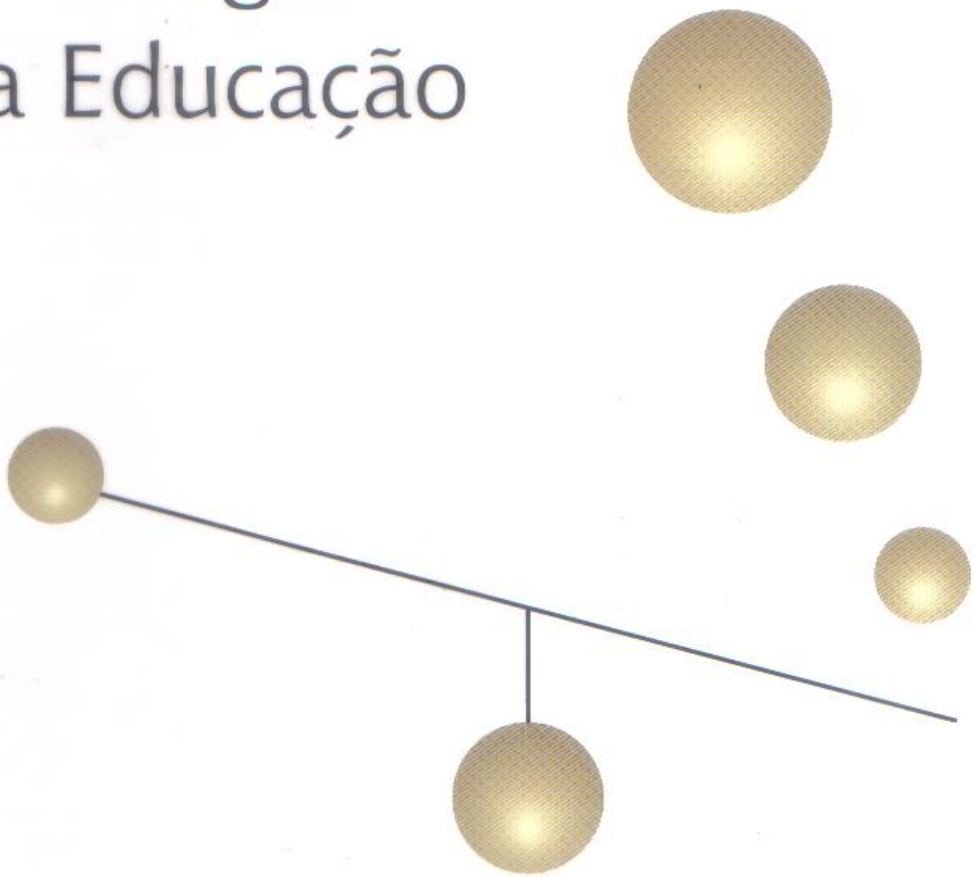
Psicometria

Teoria dos testes na Psicologia e na Educação



EDITORA VOZES

4ª Edição

L. PASQUALI

SUMÁRIO

Introdução

Desde sua primeira edição em 1998, este livro sobre Psicometria continua tendo alguns objetivos básicos que inspiraram sua confecção. Em primeiro lugar, ele visa preencher uma grave lacuna no ensino universitário na área da fundamentação epistemológica e na tecnologia de elaboração de instrumentos psicológicos de uso corrente e necessário na pesquisa e na prática profissional de ciências humanas e sociais, em especial do psicólogo e do psicopedagogo. Neste sentido, ele pretende dar as bases teóricas e de fundamentação epistemológica da medida em ciências sociais.

Em segundo lugar, o livro visa dar uma visão mais psicológica à Psicometria, procurando desvinculá-la, por um lado, da concepção tradicional baseada no materialismo científico nesta área, a qual vem sufocando o pensamento teórico dos problemas psicológicos e dar-lhe uma concepção mais cognitivista, se assim se pode dizer, e, por outro lado, dar-lhe uma visão mais psicológica e menos estatística. A Psicometria vem tradicionalmente sendo dominada por pesquisadores de cunho eminentemente estatístico. Esta situação não é um demérito para a Psicometria; pelo contrário, a estatística é fundamental neste ramo de conhecimento, sem a qual ele se torna inviável, uma vez que se trata de medir, isto é, representar o objeto psicológico via número. Ora, tratar do número utilizado na medida dos fenômenos naturais é precisamente o campo de atuação da Estatística. Entretanto, este domínio da Estatística na Psicometria fez com que esta fosse e seja concebida por muitos como um ramo da Estatística. Esta ocorrência me parece um grave erro de perspectiva, contra o qual, aliás, já nos anos 30 o próprio Thurstone, matemático e psicólogo, vinha se debatendo. A Psicometria é uma área que pretende estudar fenômenos psicológicos. Consequentemente, seu objeto específico de estudo dela são os fenômenos psicológicos e não conceitos, no caso, o número. O número, nesta ciência, é apenas o modo de representar estes fenômenos psicológicos. Assim, a Psicometria deve ser concebida como um ramo da Psicologia e que se caracteriza por expressar (observar) o fenômeno psicológico através do

número, em vez da pura descrição verbal. Nem por isso, ela deixa de ter como ponto central de sua existência o fenômeno psicológico.

Em terceiro lugar, o livro visa integrar a Psicometria clássica com a moderna, isto é, a Teoria Clássica dos Testes (TCT) e a Teoria de Resposta ao Item (TRI). Os dois enfoques possuem bases epistemológicas bastante distintas, mas nem por isso são de todo incompatíveis. O livro procura mostrar onde os enfoques coincidem e onde divergem na explicação dos instrumentos psicológicos.

Além disso, diante do receio da maioria dos cientistas sociais, em particular do psicólogo e do pedagogo, frente ao pensamento matemático, bem como o fraco preparo desses profissionais nas áreas da matemática, este livro procura enveredar o mínimo na sofisticação matemática e estatística que a Psicometria pode assumir. Obviamente, não é possível escapar totalmente do pensar matemático nesta área da medida, pois seria utilizar o número, o objeto específico das matemáticas, sem fazer uso dos princípios e métodos das mesmas. Quem entende e trabalha o número são, necessariamente, as matemáticas. Então, não há como eliminá-las no tratamento dos dados empíricos expressos via medida. Contudo, para fazer uso inteligente dos princípios e da tecnologia psicométrica não é necessário entrar nas altas sofisticações matemáticas e estatísticas que eles permitem. Evidentemente, quem é capaz de seguir por este caminho tem maiores vantagens na intelecção da problemática psicométrica, mas um bom psicometrista, sobretudo prático, não necessita ser um exímio estatístico. Ele deve ser, sim, um exímio conhecedor da teoria psicológica.

Praticamente, o livro enfoca inicialmente o problema da medida em geral nas ciências empíricas (cap. 2) e em especial na Psicologia (cap. 3). Em seguida aborda os dois modelos da Psicometria, a TCT e a TRI (cap. 4), ressaltando a problemática psicológica envolvida, que deve se sobrepor às preocupações exclusivamente estatísticas, discutindo sobretudo as bases epistemológicas dos parâmetros básicos da Psicometria. Os demais capítulos expõem as questões tradicionais tratadas em Psicometria, tanto clássicas quanto modernas, a saber, a análise dos itens (cap. 5), a questão da validade dos testes (cap. 6), da fidedignidade dos testes (cap. 7), e das normas (cap. 8). Os capítulos 9 (Equiparação dos Escores) e 10 (Os testes psicológicos e o computador) tratam de problemas atualmente em foco no caso dos testes psicológicos. O capítulo 11 aborda uma introdução à análise fatorial, técnica que continua fundamental na questão da validação dos testes psicológicos.

CAPÍTULO 1
Origem e histórico da Psicometria

1 – Introdução

Estamos nos limiares do século XX. Em Psicologia vigoravam várias tendências epistemológicas, bastante 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), 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, 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). Creio ser esta concepção a 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 Psicologia dedicada à pesquisa dos modelos matemáticos dos processos psicológicos, mas sempre a serviço destes.

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 a origem à Psicometria e fomentou seu desenvolvimento porque lhe faltava o enfoque primordial da quantificação, que era o específico da orientação da psicologia empiricista. Da psicologia introspeccionista 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, se coadunava melhor a um estudo quantitativo, pois se pode 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 e a preocupação mais teórica da Psicometria. A primeira tendência era mais aparente entre os psicólogos cujas preocupações eram mais de caráter psicopedagógico e clínico; estes psicólogos utilizavam as provas psicológicas para 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 que utilizavam a 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 foi, 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 de 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* (teste mental).

3) A *década de Binet*: 1900. Nessa década predominaram os interesses da avaliação das aptidões humanas visando à predição na área acadêmica e na área da saúde. Embora Binet se destaque, 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 lançou 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. Vários foram os fatores que concorreram para o desenvolvimento dessa era, a saber, o teste de inteligência de Binet-Simon (1905), o artigo de Spearman sobre o fator G (1904b), a revisão do teste de Binet para os Estados Unidos (Terman, 1916) e 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*).

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 eram demasiadamente dependentes da cultura onde eram criados, não apoiando a idéia de um fator geral universal, como proposto por Spearman. Tais eventos fizeram com que os psicólogos estatísticos começassem a repensar as ideias de Spearman. De fato, Kelley quebrou com a tradição de Spearman em 1928. Esta tendência foi seguida, na Inglaterra, por Thomson (1939) e Burt (1941) e nos Estados Unidos da América, por Thurstone (1935, 1947). Este último autor é especialmente relevante nesta época, pois além de desenvolver a análise fatorial múltipla, atuou no desenvolvimento da escalonagem psicológica (1927, 1928, Thurstone & Chave, 1929), bem como por ter fundado, 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. Duas tendências opostas marcam esta época: 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), publica-

da no *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) questionando o uso 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 tendência 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 melhor sintetizar o que está ocorrendo hoje no mundo da Psicometria, arrolando as principais linhas genéricas nas quais os psicometristas vêm atuando:

a) Sistematização da Psicometria Clássica: Anastasi (1988), Crocker e Algina (1986), Thorndike (1982).

b) 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), Osterlind (1983); e funcionamento diferencial dos itens (Dorans & Holland, 1992; Green, 1994; Holland & Thayer, 1988; Holland & Wainer, 1994; Swaminathan, 1994);
- construção de itens: Brown (1983), Gronlund (1988), Mehrens & Lemann (1984), Osterlind (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 Research Methods, Instruments & Computers (1969).

2.2 – Os testes psicológicos

Os testes psicológicos que foram surgindo no final do século XIX e nas primeiras décadas do século XX 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 rastreadas aos trabalhos de Galton (1822-1911) no seu laboratório em Kensington, Inglaterra.

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

1) Preocupação psicopedagógica 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, lugar para se distinguir 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 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 Estados Unidos). 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 versavam sobre o comportamento sensorial, preocupação que espelha a origem destes psicólogos como físicos e fisiologistas. 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.

Sir Francis Galton (1883), cientista, explorador e antropometrista inglês, 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. 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 por seu 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 à procura de uniformidades e não de diferenças individuais. 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 Estados Unidos. Cattell seguiu as ideias 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 correlacionavam com o desempenho acadêmico desses alunos (Wissler, 1901).

Alfred Binet e *V. Henri* (1896), psicólogos franceses, 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 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 subsequentes, especialmente nos Estados Unidos com a tradução do seu teste por Terman (1916), inaugurando de vez a era dos testes, inclusive com a introdução do QI, sendo matematicamente representado por:

$$QI = 100 (IM / IC)$$

onde,

QI = 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*). Idade mental significava o seguinte: a criança teria aquela idade mental equivalente aos itens que uma criança de dada idade cronológica respondia corretamente, acrescendo as bonificações em meses decorrentes da resolução correta de itens soltos de idades mais avançadas.

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 maciço 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, 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.

CAPÍTULO 2
Teoria da Medida

A Psicometria assume o modelo quantitativista em Psicologia. Nos manuais de metodologia científica, particularmente nas áreas das ciências psicossociais, este tema vem tratado muito esporádica e superficialmente. Fala-se quase exclusivamente de um tema muito debatido, ou seja, as escalas de números: nominal, ordinal, de intervalo e de razão. Inclusive sem demonstrar como tais escalas surgem. A problemática da medida em ciências é bem mais complexa do que isto. Para oferecer uma visão mais coerente, abrangente e racional a esta temática, o presente capítulo procura dar e explicitar os fundamentos epistemológicos de medida em ciências (empíricas), bem como explicitar os tipos e níveis possíveis que ela pode assumir.

1 – Introdução: Ciência e Matemática

A medida em ciências psicossociais, notadamente na Psicologia, deveria ser chamada puramente de Psicometria similarmente ao que ocorre em áreas afins, onde se fala de sociometria, econometria, politicometria, etc. Psicometria, contudo, tem sido abusivamente utilizada dentro de um contexto muito restrito, referindo-se a testes psicológicos e escalas psicométricas. De qualquer forma, a Psicometria ou medida em Psicologia se insere dentro da teoria da medida em geral que, por sua vez, desenvolve uma discussão epistemológica em torno da utilização do símbolo matemático (o número) no estudo científico dos fenômenos naturais. Trata-se, portanto, de uma sobreposição, ou melhor, de uma interface, entre sistemas teóricos de saber diferentes, tendo a teoria da medida a função de justificar e explicar o sentido que tal interface possui.

A matemática e a ciência empírica são sistemas teóricos[1] (ou de conhecimento) muito distintos e, em termos estruturais, não são comensuráveis. Na verdade, os dois sistemas têm objetos e metodologias próprias, distintas e irreversíveis entre si. Pode-se discernir esta distinção, atentando

---

1. Uma discussão epistemológica mais detalhada sobre sistemas de saber se encontra em Pasquali, L. (org.), Delineamento de Pesquisa em Ciência.

para a tabela 2-1. Observa-se que, em nenhum momento ou sob nenhum critério, os dois sistemas se assemelham estruturalmente. A ciência tem como referente ou objeto os fenômenos da realidade, ao passo que a matemática estuda como seu objeto o símbolo numérico (que é um conceito e não uma realidade empírica e nem uma propriedade desta realidade – Frege, 1884); a metodologia da ciência é a observação sistemática e a da matemática é a dedução; o critério de verdade para a ciência é o teste empírico, ao passo que para a matemática é a consistência interna do argumento.

Tabela 2-1. Enfoque epistemológico em ciência e matemática

Sistema Teórico	Objeto	Atitude	Metodologia	Verdade	Certeza	Critério de Verdade
Ciência (empírica)	Fenômenos naturais	Empírica	Observação e Controle	Fato	Relativa	Teste Empírico
Matemática	Símbolo numérico	Transcendental	Dedução	Teorema	Absoluta	Consistência interna do argumento

Assim, a primeira afirmação, no contexto da teoria da medida, consiste em dizer que o sistema científico do conhecimento não tem nada a ver com a matemática e vice-versa, falando-se em termos das estruturas epistemológicas dos dois saberes. O mesmo tipo de argumentação pode ser feito da ciência com relação aos outros sistemas de saber (filosofia, teologia, etc.).

2 – A natureza da medida

Apesar dessa distância epistemológica entre ciência e matemática, a primeira se apercebeu das vantagens consideráveis que ela pode obter ao se utilizar a linguagem da matemática para descrever o seu objeto próprio de estudo. Na verdade, se o modelo matemático não dita e nem fundamenta o conhecimento científico, parece que é o uso deste modelo que vem possibilitando distinguir níveis de progresso no conhecimento científico. Esta afirmação, pelo menos, aparece claramente demonstrada na ciência da Física que, com o uso do modelo matemático, pôde passar de um estágio pré-científico à física moderna. Além disso, "Os instrumentos

e técnicas de medida propiciam a ponte mais útil entre os mundos do dia a dia do leigo e dos especialistas em ciência" (Klein, 1974: 24).

O uso do número na descrição dos fenômenos naturais constitui o objeto da teoria da medida. Esta teoria está razoavelmente axiomatizada somente nas ciências físicas, aparecendo ainda lacunar nas ciências psicossociais, onde, aliás, ainda se discute a viabilidade epistemológica da própria medida.

A natureza da medida implica em alguns problemas básicos, dentre os quais três devem ser mencionados (Luce & Suppes, 1986; Suppes & Zinnes, 1963; Campbell, 1928, 1938): a representação, a unicidade e o erro.

2.1 – O problema da representação ou o isomorfismo

O problema central da medida consiste em justificar a legitimidade de se passar de procedimentos e operações empíricos (a observação) para uma representação numérica destes procedimentos. É justificável designar ou expressar objetos ou fenômenos naturais através de números? Sim, se nesta designação se preservarem tanto as propriedades estruturais do número quanto as características próprias dos atributos dos fenômenos empíricos. Trata-se do teorema da representação, isto é, representar com números (objeto da matemática) as propriedades dos fenômenos naturais (objeto da ciência).

2.2 – O problema da unicidade da representação

A questão envolvida aqui é a seguinte: será que o número é a única ou a melhor representação das propriedades dos objetos naturais para fins de conhecê-los pelo homem? Evidentemente, você vê que a resposta a esta pergunta gera imediatamente guerra, particularmente entre cientistas da área psicossocial. Entretanto, os defensores da medida em ciência respondem afirmativamente à pergunta, sem pestanejar. Mesmo assim, alertam que esta representação, ainda que sendo a melhor, apresenta níveis diferentes de qualidade ou precisão, dependendo do tipo de característica dos objetos que se está focalizando. Assim, para o caso do peso, o número representa excelente informação (a saber, o quilograma), enquanto no caso da inteligência, ele já é menos preciso (o QI, por exemplo, ou um escore num teste X). Esta problemática da unicidade da representação e de seus níveis gera os níveis da escala de medida, ou seja, define se a escala obtida será ordinal, intervalar, etc., como veremos.

2.3 – O problema do erro

A observação dos fenômenos empíricos é sempre sujeita a erros devidos tanto ao instrumental de observação (os sentidos e suas extensões através de instrumentos tecnológicos), quanto a diferenças individuais do observador, além de erros aleatórios, sem causas identificáveis. Assim, tipicamente toda e qualquer medida vem acompanhada de erros e, por consequência, o número que descreve um fenômeno empírico deve vir acompanhado de algum indicador do erro provável, o qual será analisado dentro de teorias estatísticas para determinar se o valor encontrado e que descreve o atributo empírico está dentro dos limites de aceitabilidade de medida. Note que o número matemático é um conceito unívoco, sem a mínima variabilidade de interpretação; os números são conceitos pontuais, onde um número não apresenta nenhuma interseção com o próximo. Assim, 1 é somente 1, 2 é somente 2, etc., coisa que não ocorre quando o número é utilizado para descrever (representar) fenômenos naturais, porque aqui o número 1 pode ser *mais ou menos* 1 e, assim, pode ter interseção com o 2, etc. Acontece que, na medida dos fenômenos naturais, o número se adultera um pouco, perdendo sua identidade pontual e absoluta, para se tornar um intervalo em vez de ser um ponto sem dimensões. O fato de o número, na medida, se tornar um intervalo diz que ele já tem variabilidade (variância) e isto é o erro. Se você chamar o número da matemática como número matemático, o número da medida você chamaria de número estatístico. Aquele é um ponto, enquanto este é um intervalo. Desta forma, em matemática o número está sempre solitário, inconfundível, enquanto na medida ele vem sempre acompanhado de um "cão de guarda", a variância, que indica o erro. Voltaremos a este ponto mais adiante.

Concluindo: O uso do número na descrição dos fenômenos naturais (isto é, a medida) somente se justifica se se puder responder afirmativamente às duas questões seguintes:

1 – É *legítimo* utilizar o número para descrever os fenômenos da ciência?

2 – É útil, *vantajoso*, utilizar o número para descrever os fenômenos da ciência?

O restante deste capítulo procura responder e fundamentar a resposta afirmativa a essas duas questões.

3 – A base axiomática da medida

Esta parte visa fundamentar a legitimidade epistemológica da medida em ciências, isto é, a legitimidade do uso do número como descritor de fenômenos naturais.

Há legitimidade no uso do número na descrição dos fenômenos naturais se, e somente se, as propriedades estruturais, tanto do número quanto dos fenômenos naturais, forem salvaguardadas neste procedimento. Isto é, deverá haver isomorfismo estrito (relação de 1 para 1) entre propriedades do número e aspectos dos atributos da realidade empírica.

São propriedades básicas do sistema numérico: a identidade, a ordem e a aditividade. A medida deve resguardar, pelo menos, as duas primeiras destas propriedades; de preferência, as três.

Para melhor enquadrar a Psicometria e a medida em geral em ciências psicossociais, a base axiomática da medida será tratada dentro das ciências físicas, fazendo em seguida as ressalvas e correções necessárias para o caso da medida em ciências psicossociais e, em especial, da Psicologia.

3.1 – Axiomas do sistema numérico

Stevens (1951) representaria o sistema numérico como na Figura 2-1.

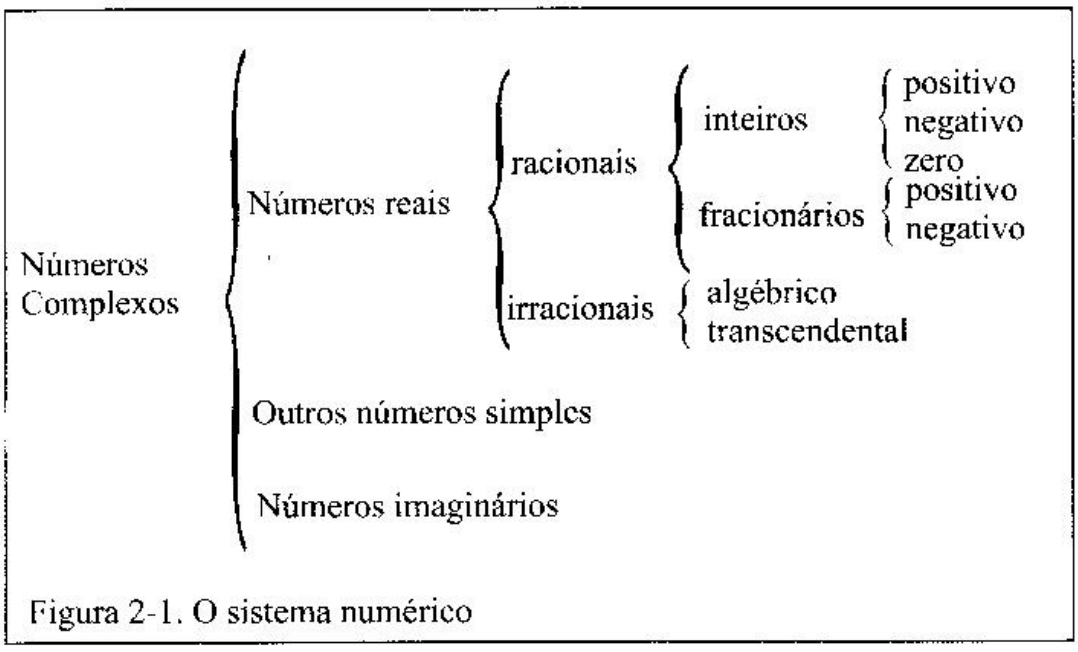


Figura 2-1. O sistema numérico

Estes vários tipos de números surgiram em épocas históricas diferentes, segundo as necessidades dos estudiosos e as da vida prática. Inicialmente só havia os números inteiros, que se mostravam suficientes para a contagem de objetos discretos; razão pela qual também são chamados de números naturais (cf. Klein, 1974). Com eles se podia fazer as operações de soma e de multiplicação. Eles não davam sempre certo, especialmente quando se queria subtrair um número maior de um número menor. Esta limitação do sistema de inteiros fez com que o sistema fosse estendido para incluir números negativos e o zero. Com a divisão, o sistema de inteiros se mostrava ainda mais limitado, o que forçou a adoção de números fracionários. Este conjunto de números (inteiros positivos, negativos, zero e frações) constitui o sistema de números racionais, dado que qualquer número deste sistema pode ser expresso em termos de razão entre dois números inteiros. Excetuada a divisão por zero, todas as operações são possíveis dentro deste sistema numérico. Contudo, certas operações matemáticas não eram viáveis dentro do sistema, como, por exemplo, a raiz quadrada de 2. Inventaram-se, então, os números irracionais, e assim se fechou o círculo dos números reais, suficientes para permitir qualquer sorte de medida da realidade, até o presente.

A matemática é um saber baseado em puras convenções; assim, tanto o seu objeto (o número) quanto suas regras são convencionadas. No início do século XX, os filósofos e matemáticos Whitehead e Russell (1910-1913; 1965) elencaram nada menos que 27 axiomas ou regras do jogo da matemática, apresentadas em seu livro *Principia Mathematica*. Destes axiomas, três grandes conjuntos são importantes para o caso da medida. Trata-se dos axiomas que definem as propriedades numéricas de identidade, ordem e aditividade.

1) *Identidade*. Esta propriedade define o conceito de igualdade, isto é, que um número é idêntico a si mesmo e somente a si mesmo. Ela apresenta três axiomas (postulados aceitos e não provados) que expressam a relação de IGUAL A (=):

- reflexividade: $a = a$ ou $a \neq b$. Números são idênticos ou são diferentes;
- simetria: se $a = b$, então $b = a$;
- transitividade: se $a = b$ e $b = c$, então $a = c$. Duas coisas iguais a uma terceira são iguais entre si.

2) *Ordem*. Esta propriedade se baseia na desigualdade dos números. Todo número é diferente do outro. Essa desigualdade não é somente de qualidade, mas ela se caracteriza em termos de magnitude, isto é, um número não é somente diferente do outro, mas um é maior que o outro. Aliás, eles são diferentes precisamente porque um é maior que o outro. Assim, excetuado o caso de igualdade, os números podem ser colocados numa sequência invariável ao longo de uma escala linear: sequência monotônica crescente. Também apresenta três axiomas, que expressam NÃO IGUAL A ou MAIOR QUE (>):

– assimetria: se a > b, então b ≠ a. A ordem dos termos não pode ser invertida;

– transitividade: se a > b e b > c, então a > c;

– conectividade: ou a > b ou b > a;

– um quarto axioma é o de ordem-denso: números racionais são tais que entre dois números inteiros quaisquer há sempre um número racional; o intervalo entre dois inteiros não é vazio.

3) *Aditividade*. Os números podem ser somados, isto é, podem ser concatenados de modo que a soma de dois números, excetuado o zero, produz um outro número diferente deles próprios. Isto é, as quatro operações (soma, subtração, multiplicação e divisão – as três últimas são redutíveis à primeira) podem ser aplicadas aos números. Dois axiomas se destacam:

– comutatividade: a + b = b + a. A ordem dos termos não altera o resultado da adição;

– associatividade: (a + b) + c é igual a a + (b + c). A ordem de associação ou de combinação dos termos não afeta o resultado.

Nota: As quatro operações básicas dos números (soma, subtração, multiplicação, divisão) se reduzem à soma. Veja:

Soma:	3 + 2 = 5
Subtração:	3 – 2 = 3 + (-2)
Multiplicação:	3 x 2 = 2 + 2 + 2 (três vezes o 2)
Divisão:	4 ÷ 2 = ½ + ½ + ½ + ½ (quatro vezes ½).

3.2 – Axiomas da medida

Como a medida consiste na atribuição de números às propriedades das coisas segundo certas regras, ela deve garantir que as operações empíricas salvem os axiomas dos números. A medida que salva todos esses axiomas é a mais sofisticada possível e, por isso, rara (escala de razão). A maioria das medidas, ao menos em ciências psicossociais, se dão por satisfeitas se puderem salvar, pelo menos, os axiomas de ordem. Se somente os axiomas de identidade forem salvos (escala nominal), a operação propriamente não chega a ser medida, mas trata-se apenas de classificação, pois a única característica do número salva é a sua identidade; isto é, o número utilizado para uma operação empírica deve ser diferente do de uma outra operação. Para tanto, aliás, o número é utilizado tão somente como numeral, a saber, um rabisco diferente de outro, que poderia ser substituído por qualquer outro sinal ou rabisco (desde que diferentes entre si) sem a menor consequência para a medida. O número neste caso serve apenas de etiqueta para uma classe de coisas. A medida realmente acontece quando se salvam, pelo menos, os axiomas de ordem dos números. Então fica a pergunta: é possível se demonstrar a existência de ordem de magnitude nos atributos das coisas, isto é, as coisas têm dimensões? (entendidas estas como atributos mensuráveis, propriedades empíricas possuidoras de magnitude). Como resposta a esta questão, poder-se-ia simplesmente assumir que sim: os atributos empíricos têm magnitude, como o senso comum nos parece dizer quotidianamente quando fala de 'mais do que', 'maior que' e expressões similares. Contudo, esta não parece ser uma base suficientemente segura para fundamentar uma teoria da medida. É preciso, então, demonstrar empiricamente que tal ocorrência existe na realidade das coisas. Nas ciências físicas esta questão parece resolvida, mas nas ciências psicossociais ainda suscita acirradas controvérsias. Antes de oferecer uma tentativa de demonstração experimental de axiomas da medida, note o comentário a seguir.

Comentário: Os números, pelo menos os complexos, diferem entre si em termos de magnitude e não em termos de qualidade. Assim, o 2 não é qualitativamente diferente do 1; ele é diferente quantitativamente. Quer dizer que todos os números são da mesma qualidade ou natureza, isto é, são todos da qualidade de quantidade ou de magnitude; eles diferem em ordem, porque um é maior que o outro. Repare que os fenômenos naturais, por outro lado, diferem entre si também qualitativamente: a cor é uma qualidade diferente da do peso, que é diferente da do comprimento, etc.

Estas diferenças são de qualidade, pois se trata de aspectos, qualidades, características, atributos [...] diversos e distintos de um mesmo objeto natural. O isomoformismo defendido na medida é entre as propriedades do número (sobretudo de ordem e aditividade) com estas mesmas propriedades de cada atributo de um objeto natural e não com o próprio objeto natural. De fato, os diferentes atributos de um mesmo objeto não diferem em termos de quantidade e sim de qualidade. Entretanto, um mesmo atributo de um objeto natural pode variar em diferentes momentos ou situações ou objetos-indivíduos e esta variação é que é definida em termos de magnitude, portanto mensurável (donde a medida: esta é de atributos individuais de objetos e não dos próprios objetos). Assim, por exemplo, da rosa eu posso falar de diferentes qualidades, tais como intensidade de aroma, peso, tamanho, etc. Mas cada uma dessas qualidades pode aparecer com magnitude diferente de aroma, peso, comprimento, etc. (isto é, eu posso medir as magnitudes de cada qualidade). Em termos de seus atributos, qualitativamente diferentes, não posso dizer que uma rosa é mais rosa que a outra; posso, porém, dizer que uma é mais aromática, mais pesada, etc. que a outra. Cf. a figura 2-2 para facilitar a compreensão.

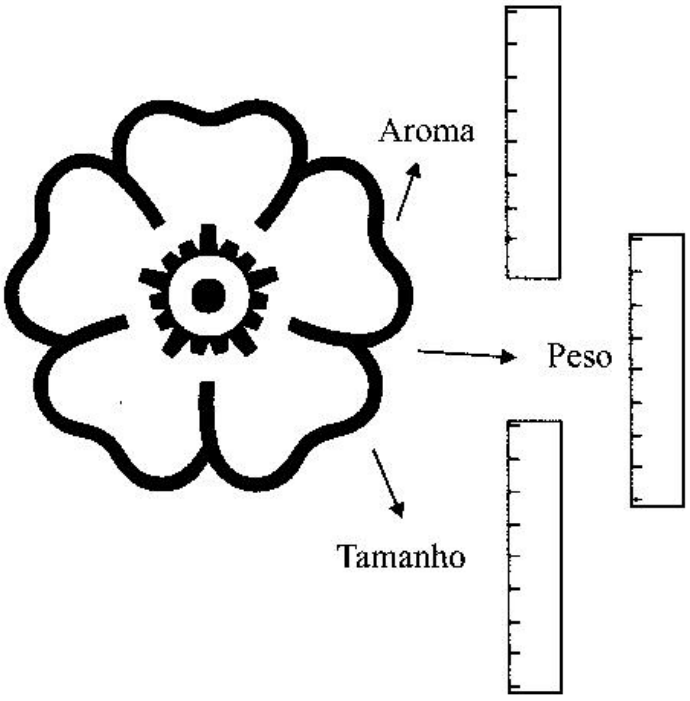


Figura 2-2. Três qualidades da rosa com suas magnitudes

3.2.1 – Demonstração empírica dos axiomas de ordem (Guilford, 1954)

Os axiomas de ordem afirmam que, na medida, a ordem dada pelos números atribuídos aos objetos (transitividade e conectividade) deve ser a mesma obtida pela ordenação empírica destes mesmos objetos. Existe ordem ("maior que") nas propriedades das coisas. Exemplos: Um metal que arranha um outro e não pode ser arranhado por este, diz-se que é mais duro. Assim, uma ordem empírica de dureza pode ser estabelecida a partir da operação empírica de arranhar. Igualmente, o alinhamento de linhas mostra que uma é maior que outra, donde uma ordenação de objetos em termos de comprimento poder ser montada. Um tom que é dito mais alto que outro por uma amostra de sujeitos, diz-se que é mais agudo. Assim, uma ordem de altura tonal (*pitch*) pode ser estabelecida. Idem, se um sujeito resolve corretamente maior número de uma série de problemas do que outro, diz-se que é mais inteligente. Assim, pode-se estabelecer uma escala de inteligência. As inversões que ocorrem são consideradas "erros de medida" ou de observação, que devem ser tratados dentro da teoria da consistência, a qual visa mostrar que, apesar desses erros, há consistência na medida.

3.2.2 – Demonstração empírica dos *axiomas de aditividade*

A demonstração dos axiomas de aditividade parece ser possível somente no caso dos atributos extensivos, como massa, comprimento e duração temporal, bem como no caso da probabilidade. A aditividade se baseia na ideia de concatenação: a combinação (concatenação) de dois objetos ou eventos produz um terceiro objeto ou evento com as mesmas propriedades dos dois, mas em grau maior. Assim, tomando-se um objeto de comprimento 'x' (medido em uma unidade de comprimento qualquer, o metro, por exemplo), encontrar um outro objeto com o mesmo comprimento 'x', juntando (concatenando) os dois objetos obtém-se um objeto maior 'z' com comprimento duas vezes os comprimentos dos objetos individuais. O conceito de concatenação implica que A com B (A concatenando B) = A + B.

Conclusão: Fica, assim, claro e amplamente demonstrado que a medida, isto é, a utilização do número para descrever os fenômenos naturais, é legítima e adequada. Contudo, da discussão anterior, você também percebe que existe medir e medir; ou seja, nem todas as medidas são iguais, digamos, em qualidade. Esta qualidade depende do grau ou da quantidade de isomorfismo que existe entre as propriedades do número e

as propriedades dos fenômenos naturais. Isto quer dizer que há níveis diferentes de correspondência entre o número e os fenômenos naturais, o que implicará diferentes níveis de medida, como veremos em seguida.

4 – Níveis da medida (Escalas de medida)

Dependendo da quantidade de axiomas do número que a medida salva, resultam vários níveis de medida, conhecidos como escalas de medida. Como vimos, 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 se aproximará 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. 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 a seguir ilustra como se originam as várias escalas de medida:

		Origem	
		Não Natural	Natural
Intervalo	Não Igual	Ordinal	Ordinal
	Igual	Intervalar	Razão

Assim, uma medida de uma propriedade de um objeto natural que não tem uma origem natural (exemplos: aroma, QI, amizade) não pode começar em zero (0), porque não se conhece um valor zero de aroma ou de QI. Mesmo se você usa o zero na medida de tais atributos, este é um zero fictício, não natural. Desta forma, uma escala de medida de tais atributos pode começar com qualquer número, inclusive o zero, sendo este número a origem da escala e o próximo número tem que ser maior (se a escala for ascendente) porque a escala precisa salvar, pelo menos, a ordem natural dos números. Uma tal escala seria chamada de ordinal, onde a origem é arbitrária e a distância entre os números não seria igual. Consequentemente, as seguintes escalas são equivalentes, produzem exatamente a mesma informação:

3	4	5	6	7
3	5	6	10	100
0	1	2	3	4
-3	0	15	30	31

A única coisa relevante que distingue estas escalas é uma questão de estética, sendo provavelmente a mais elegante a escala 0 1 2 3 4. Mas elegância é questão de gosto e "de gustibus non est disputandum" (não se briga por gostos). Agora, seria um erro transformar estas escalas na seguinte:

0	1	2	4	3

porque se perderia a ordem (monotônica crescente).

Se nesta mesma medida você puder salvar a origem natural, isto é, o zero, mas não puder salvar o intervalo igual entre os números da escala, você ainda estaria medindo apenas ao nível ordinal. Por exemplo: medir o peso de diferentes objetos sem ter uma balança. Neste caso, você pode pedir a um ou vários sujeitos para ordenar os objetos em termos de mais pesado, surgindo daí uma ordenação dos mesmos pelo peso sem se poder dizer quanto um objeto é mais pesado que o anterior. Peso, na verdade, é um atributo da natureza que permite o valor 0, mas o processo de medida, como descrito, não permite dizer mais do que um objeto ser mais pesado

que o outro, sem se poder definir quanto mais. Se você pudesse ou puder definir quanto mais pesado ele é, então você já estaria medindo ao nível de escala de razão que, além de ter uma origem natural 0, tem intervalos iguais entre os números da escala. Uma tal escala sempre começa com 0 e seus números estão a distâncias iguais entre si. Exemplo:

0	1	2	3	4
0	2	4	6	8
0	5	10	15	20

Nesta escala o que muda é apenas a unidade de medida (o tamanho do intervalo), sendo ela sucessivamente de 1, 2 e 5, no exemplo proposto.

Um exemplo de uma escala simplesmente intervalar e suas transformações legítimas seria a seguinte:

0	2	4	6	8
2	4	6	8	10
-5	0	5	10	15

onde são salvos a ordem dos números e o tamanho do intervalo entre eles.

Note que o fator que define o nível da medida não é o número, mas sim a característica do atributo medido da natureza (da realidade): se ele permite ou não uma ordem natural, o 0 (tais como peso, comprimento....), se permite definir distâncias iguais ou não (muitos pesquisadores afirmam que nenhum atributo não extensivo da natureza, como todos os atributos psicossociais, permite medida intervalar!). Os números, por natureza, têm todas estas características: origem natural (o 0), ordem e distâncias iguais entre si. Assim, para os números, todas as escalas são de razão, mas para a medida, isto é, os números utilizados para descrever fenômenos naturais, nem sempre se pode salvar estas características dos números.

A tabela 2-2 sumaria as características de cada escala.

Tabela 2-2. 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)	Frequências: f, %, p, Mo, qui2, C
Ordinal	identidade – ordem	– ordem	intervalo – origem – unidade	Monotônica crescente (isotonia)	Não paramétricas: Md, r_S, 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

f = frequência; % = percentagem; p = proporção; C = coeficiente de contingência; Md = mediana; DP = desvio padrão; r_s = correlação de Spearman; U = teste de Mann-Whitney; r = correlação produto-momento de Pearson; F = teste de Fisher (análise da variância)

Como já insinuado, 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. Por exemplo, posso organizar o leque de idades dos sujeitos em quatro grupos etários (adolescência, jovem-adulto, adulto e terceira idade); mas, nesse caso, a partir desses grupos não posso saber a média das idades da amostra. 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 utilizarem procedimentos estatísticos superiores. São chamados 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.

5 – Formas e unidades de medida

Até aqui sabemos que posso expressar legitimamente os atributos naturais com escalas de números e que estas se apresentam em diferentes níveis. Fica, entretanto, o seguinte problema: como é que vou atribuir um número a tal e tal atributo de um fenômeno natural e por que este número e não outro? Isto não pode se constituir em um processo aleatório, pois medir (como veremos mais adiante) visa precisamente dar maior precisão à descrição do fenômeno natural e nada é mais impreciso que uma descrição aleatória. Então, o que fazer? Bem, se cada atributo da realidade empírica apresentasse uma unidade-base natural específica de magnitude, a medida dele seria uma tarefa relativamente fácil. Seria suficiente verificar quantas unidades-base ele possui e o número de unidades seria a medida do atributo em questão. Desta forma, se eu pudesse dividir um dado atributo em pedacinhos, o tamanho deste atributo seria a soma desses pedacinhos. Para tal tarefa, eu posso definir um pedacinho qualquer aleatoriamente, como por exemplo o centímetro para o comprimento, e assim ver quantos deles este atributo tem e quanto aquele... Acontece, porém, que nem no mundo da física todos os atributos permitem uma definição de unidade-base natural específica, como por exemplo é o caso da velocidade. Disto resulta que deve haver mais de uma forma de se proceder à medida dos atributos da realidade que não seja a simples enumeração do número de unidades que o objeto apresenta.

5.1 – Formas de medida

Há diferentes maneiras (formas) de se atribuir números às propriedades dos objetos. Uma das taxonomias mais úteis consiste em distinguir três formas diferentes de mensuração: medida fundamental, medida derivada e medida por teoria (esta chamada de medida "by fiat" por Campbell, 1928, 1938). Pode-se igualmente falar em medida direta e medida indireta; e há outras ainda. A primeira, contudo, parece mais esclarecedora.

5.1.1 – Medida fundamental

É a medida de atributos de objetos empíricos para os quais, além de se poder estabelecer uma unidade-base natural específica, existe uma representação extensiva. São dimensões (atributos mensuráveis) que permitem a concatenação, isto é, dois objetos podem ser associados, concatenados, for-

mando um terceiro objeto de mesma natureza. Tal situação ocorre com os atributos de massa, comprimento e duração temporal. Estes atributos permitem uma medida direta e fundamental, dado que o instrumento utilizado para medi-los possui a mesma qualidade que se quer medir neles. Assim, ao se medir o comprimento de um objeto, utiliza-se um instrumento composto de unidades de comprimento. A medida dele será dada pela coincidência de pontos entre o comprimento do objeto e a unidade de comprimento marcada no instrumento, por exemplo o metro. É como se você dissesse: este objeto tem 100 centímetros de comprimento, isto é, tem o tamanho da soma de 100 pedacinhos de comprimento, sendo estes pedacinhos os centímetros. O metro é um analógico conveniente composto de 100 destes pedacinhos ou centímetros (veremos em seguida que hoje em dia se define o metro com outro tipo de pedacinhos de comprimento) que pode ser utilizado para facilitar a medida do comprimento das coisas.

Mesmo podendo ser possível conceitualmente se proceder a uma medida fundamental nos casos mencionados, nem sempre isto é empiricamente factível. Por exemplo, como se faria uma medida fundamental de distâncias astronômicas ou subatômicas? Ou como se poderia medir fundamentalmente a massa de uma galáxia? Nestes casos e semelhantes é preciso recorrer a outras estratégias de medida, mais indiretas, como a medida derivada ou outra.

5.1.2 – Medida derivada

Muitos atributos da realidade não permitem uma medida extensiva ou possuem unidades-base e, portanto, nenhuma medida fundamental é deles possível. Podem, contudo, ser medidos indiretamente através do estabelecimento de uma relação com medidas extensivas. Este procedimento depende da prova empírica de que estes atributos são afetados independentemente por dois ou mais componentes. Se estes componentes permitem medida fundamental, então se pode obter uma medida derivada para aqueles atributos não extensivos através de uma função de potência entre os componentes do qual o atributo em questão é constituído. De qualquer forma, uma tal medida é derivada se finalmente ela pode ser expressa em termos de medidas fundamentais. Por exemplo, sabe-se que a massa varia em função de volume e de densidade: massa = volume x densidade. Como a massa permite medida fundamental (peso, expresso em quilos) e o volume também (o cubo do comprimento = m^3), então a densidade, que não possui medida fun-

damental, pode ser medida indiretamente em função de massa e volume (quilos dividido por metros cúbicos = kg/m^3).

Deve-se notar que o fundamento da função existente entre os componentes constitui uma lei, isto é, deve ser um dado empiricamente demonstrado e não somente baseado em alguma teoria. Assim, a massa sendo determinada pelo volume e pela densidade é uma descoberta científica, uma lei, não uma hipótese. Entende-se, portanto, por medida derivada de um atributo aquela cujos componentes do atributo, estabelecidos por uma lei empírica, tenham finalmente dimensões extensivas.

Nessa discussão sobre as formas de medida estamos falando de atributos extensivos ou atributos possuidores de unidades-base como sendo sinônimos. Na verdade, esta sinonímia só entrou em jogo em 1960 com a definição do conceito de unidade-base e do estabelecimento das primeiras unidades-base da Física do *Système International des Unités* – SI (como veremos logo mais). Até essa data, eram em número muito reduzido as propriedades naturais consideradas extensivas; elas praticamente se esgotavam com os atributos de comprimento, peso e duração temporal. Agora, com a definição de atributo extensivo sendo aquele para o qual existe unidade-base de medida, o número de propriedades extensivas subiu para 6 (em 1960) e 7 (em 1961). Assim, por exemplo, a luminância pode ser considerada um atributo extensivo já que possui unidade-base, a saber, a candela. De fato, ela é lum = cd/m^2. O mesmo vale para resistência elétrica, força do campo elétrico e do campo magnético.

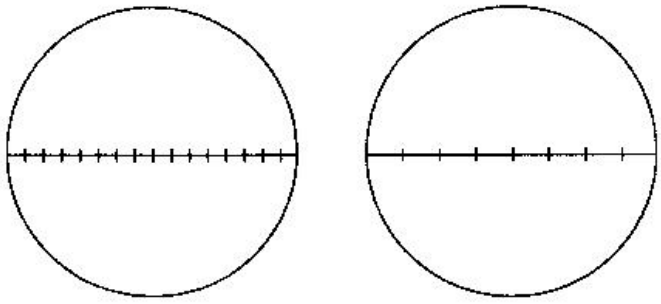
5.1.3 – Medida por teoria

Há outros atributos da realidade, e é o caso de quase todos em ciências psicossociais, que: (1) não podem ser expressos em termos de dimensões extensivas ou de unidades-base, não permitindo, consequentemente, uma medida fundamental, e (2) não são resultantes de componentes extensivos, não permitindo consequentemente nem a medida derivada. Tais atributos são mensuráveis somente com base em leis e em teorias científicas.

1) *Medida por lei*: quando uma lei for empiricamente estabelecida entre duas ou mais variáveis, a(s) constante(s) típica(s) do sistema pode(m) ser medidas indiretamente através da relação estabelecida entre estas variáveis, como é o caso da viscosidade em Física e a lei do reforço em Psicologia.

2) *Medida por teoria*: quando nem leis existem relacionando variáveis, pode-se recorrer a teorias que hipotetizam relações entre os atributos da realidade, permitindo assim a medida indireta de um atributo através de fenômenos a ele relacionados via teoria. O importante neste caso é garantir que haja instrumentos calibrados para medir (fundamentalmente ou de outra forma válida) os fenômenos com os quais o atributo em questão esteja relacionado pela teoria. Mesmo em Física isto ocorre, como é o caso da medida das distâncias galácticas. Afirma-se, por exemplo, que medindo o movimento das linhas espectrais para o vermelho estar-se-ia medindo as distâncias astronômicas, dada a teoria de que existe uma relação sistemática entre a distância de uma galáxia e a velocidade do seu afastamento e a cor do seu espectro luminoso. O mesmo vale para o efeito Doppler que afirma (teoria) que um objeto que se afasta tende a espalhar as ondas do seu espectro luminoso, reduzindo sua frequência.

Veja, por exemplo, como funciona a teoria do efeito Doppler. Este cientista alemão hipotetizou que quanto mais as linhas do espectro luminoso da galáxia se separam, mais rapidamente esta está se afastando, e quanto mais elas se aproximam, mais rapidamente a galáxia está se aproximando de nós. Veja a ilustração da separação das linhas espectrais:



Aproximação Afastamento

Há duas questões importantes nesta hipótese de Doppler:

1) é preciso uma medida extremamente exata das distâncias entre as linhas espectrais; isto é teoricamente de simples solução, pois se trata de uma medida fundamental de comprimento (distância) que o metro resolve. Ela é, contudo, um problema tecnicamente difícil, porque nesta medida, como em qualquer medida, existe sempre o erro e um erro mínimo aqui irá resultar em erro gigantesco ao se referenciar às distâncias intergalácticas;

2) outra questão, e esta é a mais importante, é a seguinte: é verdade que a distância das linhas espectrais tem a ver com o movimento da galáxia? Na resposta a esta questão é onde entra a teoria dizendo que sim. Se isto for verdade, então posso deduzir daí hipóteses sobre o movimento das galáxias e testá-las empiricamente e tornar, assim, a teoria uma teoria científica. Foi o que os físicos fizeram com sucesso, tornando a medida por teoria uma medida cientificamente legítima.

5.2 – Unidades de medida

Normalmente existe interdependência entre os fenômenos, de sorte que, ao se variar um deles, o outro covaria com ele. Esta covariância pode ser expressa por alguma constante. Estas constantes podem ser universais, como o caso da gravitação universal que covaria com as gravitações locais de um sistema menor, por exemplo, a da massa, chamada aquela inclusive de constante universal de gravitação. Outras constantes pertencem a algum sistema específico, chamadas de constantes do sistema ou locais, como a constante entre massa e volume ou as constantes da lei do reforço em Psicologia. Evidentemente, a descrição de tais constantes pode constituir uma medida indireta.

Além de constantes que relacionam dois ou mais atributos, os próprios atributos variam por conta própria, assumindo diferentes magnitudes, isto é, eles são dimensões, entendendo por isso que podem variar de magnitude e, portanto, podem ser mensuráveis. Neste caso, seria extremamente útil se houvesse, para cada atributo diferente, uma unidade básica com a qual se pudesse determinar a magnitude do mesmo. De fato, qualquer unidade que se queira definir serve aos propósitos da medida, bastando haver consenso sobre a mesma. Mas é fácil ver as vantagens de se estabelecerem unidades-base aceitáveis para todos. Nas ciências físicas, este esforço tem sido constante. O critério que tem guiado os físicos na procura destas unidades-base foi a busca de um fenômeno natural de estabilidade máxima que pudesse servir como padrão físico da unidade-base para o sistema. A história da procura destas unidades tem lances de Babel, pois cada região do mundo tinha seus sistemas de medida, incomensuráveis com outras regiões. Por exemplo, para medir o comprimento na França se utilizava o pé do rei francês, que era diferente do tamanho do pé direito do rei da Inglaterra, e daí surgiam brigas sem fim no comércio entre os países... Há cerca de 200 anos, contudo, uma procura mais sistemá-

tica e mais entrosada no âmbito mundial tem sido desenvolvida até que culminasse no *Système International des Unités* (abreviado SI), definido na *11th General Conference On Weights and Measures* (Paris, 1960), onde foram estabelecidas seis unidades-base ou primárias para os fenômenos físicos, sendo todas as restantes medidas derivadas destas seis primárias (Klein, 1974; Luce & Suppes, 1986). No ano seguinte, uma sétima unidade-base foi definida, o mole, que representa a substância (peso, massa) da molécula e é igual à soma dos pesos atômicos de todos os átomos que compõem a molécula. Lembrando que o peso atômico (também chamado número de massa) corresponde ao total de nucleídeos (prótons e nêutrons) do núcleo do átomo. A tabela 2-3 sintetiza estas unidades-base consensuais.

Tabela 2-3. Unidades-base da Física

Atributo	Unidade	Sigla	Padrão Físico (Definição do SI)
Comprimento	metro	m	"O metro é o comprimento igual a 1.650.763,73 comprimentos de onda no vácuo da radiação correspondente à transição entre os níveis $2\ p^{10}$ e $5\ d^5$ do átomo do Krypton-86".
Massa	quilograma	kg	"O quilograma (unidade de massa) é a massa de um cilindro especial feito de liga de platina e de irídio, que é considerado como o protótipo internacional do quilograma, e é conservado sob os cuidados do *Bureau International des Poids et Mesures* num cofre forte em Sèvres, França".
Tempo	segundo	s	"O segundo é a duração de 9.192.631.770 períodos (ou ciclos) da radiação correspondente à transição entre dois níveis hiperfinos do átomo de césio-133".
Corrente elétrica	ampère	A	"O ampère, unidade de corrente elétrica, é a corrente constante que, se mantida em dois condutores paralelos de comprimento infinito, de uma grossura negligível, e colocados a 1 metro de distância num vácuo, produzirá, entre estes condutores, uma força igual a $2x10^{-7}$ newtons por metro de comprimento (cerca de 0,1 kg)".
Temperatura	Kelvin	K	"O kelvin, a unidade de temperatura termodinâmica, é a fração 1/273,16 da temperatura termodinâmica do triplo ponto da água" (no qual gelo, água e vapor estão em equilíbrio – igual a -273,16°C)".
Intensidade da luz	candela	cd	"Luminosidade de 1/600.000 de um metro quadrado de pura platina fundida no ponto de se solidificar. Isto corresponde a uma temperatura de 2.045°K".
Massa atômica	mole	mol	Montante de substância que corresponde à soma dos pesos atômicos de todos os átomos que compõem uma molécula

A maioria das outras unidades em física é expressa em unidades derivadas destas seis unidades-base. Por exemplo, densidade é igual a peso por volume (kg/m^3), velocidade a metros por segundos (m/s), luminância a intensidade da luz por área que é expressa em termos de distância (cd/m^2), volt é watts por ampère (V = W/A), watt é joule por segundo (W = j/s), joule é newton vezes comprimento (j = N.m), newton é peso vezes distância por tempo ($N = kg \times m/s^2$), etc.

A procura de unidades similares em ciências psicossociais é algo ainda precário, exceto onde medidas fundamentais forem possíveis, como talvez em psicofísica (medida dos estímulos) e na análise experimental do comportamento (medidas de estímulos e frequência de respostas). E, por isso, nestas ciências, prevalece a forma de medida por teoria como a corriqueira.

6 – A medida em ciências psicossociais

Medidas fundamentais nestas ciências parece difícil de serem concebidas. Mesmo em economia, que se apresenta como a mais desenvolvida nesta área, parece ter caído em descrédito a concepção de que a escolha dos sujeitos se reduziria à avaliação da quantidade e preço dos bens. De fato, há ali fatores subjetivos que codeterminam a escolha dos sujeitos, fatores agrupados sob o construto de utilidade. Também não parece aceitável que a utilidade de um conjunto de bens possa ser reduzida à soma das utilidades individuais destes bens. Deste problema surgiu a teoria moderna da utilidade baseada na teoria dos jogos. Em psicofísica também se tenta enquadrar a medida como sendo fundamental. Entretanto, para isso dever-se-ia modificar a definição de medida fundamental como sendo a medida de atributos extensivos. Em psicofísica o atributo de interesse é a resposta do sujeito a estímulos físicos. Estes certamente podem permitir medida fundamental, mas não são eles que constituem o interesse específico direto da medida psicofísica, mas sim a resposta a eles. E desta não há como visualizar uma medida fundamental, dado que não é um atributo extensivo. A medida da resposta se faz em função da sua relação com o estímulo, relação estabelecida por uma lei empiricamente demonstrada. A medida, portanto, se baseia numa função entre "componentes". A palavra "componentes" está entre aspas porque o estímulo realmente não é componente da resposta no sentido dado nas medidas derivadas, nas quais os componentes relacionados são propriedades constituintes do atributo medido derivadamente, como massa em função de volume e densidade.

Se medida fundamental não é defensável em ciências psicossociais, nem a derivada o é. Resta, então, a possibilidade de se medir nestas ciências por uma terceira forma, que vimos apresentando sob a égide de medida por teoria, que congrega finalmente aquelas formas de medida não redutíveis a medidas fundamentais. Duas formas de medida são aqui destacáveis: medida por lei e medida por teoria propriamente. As duas podem ser enquadradas sob medida por teoria, dado que a lei constitui uma hipótese derivável de alguma teoria e empiricamente demonstrável.

6.1 – Medida por lei

A medida por lei é comum nas ciências psicossociais. Em Psicologia, em particular, ela faz parte da história da psicofísica e da análise experimental do comportamento. Em psicofísica, a história que vai de Weber a Stevens é a da medida por lei: lei da constante (Weber), lei logarítmica (Fechner) e lei da potência (Stevens). Na análise experimental do comportamento temos as várias leis do reforço, por exemplo.

Em que consiste uma medida por lei? Mede-se por lei quando se quer demonstrar empiricamente que dois ou mais atributos estruturalmente diferentes mantêm entre si relações sistemáticas. Duas condições são expressas nesta concepção: (1) os atributos são de natureza diferente, um não é redutível ao outro. Por exemplo: a cor e a distância são dois atributos distintos dos fenômenos físicos (no caso do desvio para o vermelho das linhas espectrais dos objetos na medida de distâncias). No caso da medida psicofísica e da análise experimental do comportamento, acontece o mesmo com a resposta e o estímulo, que são dois atributos diferentes; e (2) a existência de uma relação sistemática entre estes atributos, que foi demonstrada cientificamente. Assim, as manipulações efetuadas num atributo repercutem sistematicamente no outro, donde é possível estabelecer uma função de covariância entre os dois, uma lei.

6.2 – Medida por teoria

Uma teoria não é uma lei, dado que é composta de axiomas ou postulados e não de fatos empíricos. Ademais, ela é científica se de seus axiomas é possível deduzir hipóteses empiricamente testáveis. O caso da medida por teoria ocorre também em Física, como ficou dito acima. No caso da Psicologia, podemos distinguir vários enfoques teóricos com respeito à medida por teoria. Quatro deles são de uso corrente, quais sejam:

1) *Teoria dos Jogos*: esta trabalha basicamente com dois parâmetros, isto é, (a) a probabilidade objetiva de ganhos e perdas associada com a escolha de cada alternativa disponível e (b) a utilidade, que expressa a preferência subjetiva do sujeito por uma determinada alternativa. O conceito de utilidade foi introduzido pela ciência econômica, diante do fato de que os sujeitos nem sempre escolhem a alternativa de maior probabilidade objetiva de ganhos. De fato, as alternativas numa dada situação podem ser ordenadas tanto pela grandeza de suas probabilidades objetivas quanto pelo nível de preferência que o sujeito lhes atribui. Aliás, o conceito de utilidade, entendida como a força de nosso desejo, já foi trabalhado por Pascal em *La logique*, ou *l'art de penser* (1662, apud Bernstein, 1997), depois por Bernoulli e redescoberta por Bentham no século XVIII; este conceito foi reelaborado por Von Neumann e Morgenstern na teoria dos jogos, que publicaram em *Theory of games and economic behavior* (1953). A escolha ou decisão final depende de uma interação entre estas duas ordenações, decisão que a teoria dos jogos procura explicar (Von Neumann & Morgenstern, 1944; Blinder, 1982; Zagare, 1984; Mirowski, 1991, 1992; Macrae, 1992; Nasar, 1994; Leonard, 1995).

2) *Teoria Psicofísica*: esta teoria trabalha com estímulos e respostas. Dentro dela se distinguem: (1) *Teoria clássica* que trabalha sobretudo o problema dos limiares sensoriais (Weber, 1834; Fechner, 1860; Gescheider, 1997); (2) *Teoria da Detecção do Sinal*: esta trabalha com dois parâmetros, a saber, relação sinal-ruído ('d') e a disposição do sujeito ('beta'). O primeiro parâmetro define o grau de detectabilidade do sinal contra um fundo de ruído e o beta define o nível de vontade ou disposição que o sujeito tem de ver o sinal quando ele está presente (Gescheider, 1997; Swets, 1959; Green & Swets 1966; Swets et al., 1961); (3) *Teoria stevensiana* (Stevens, 1946, 1951, 1959, 1960, 1971, 1974, 1975; Faleiros Souza, Kamizaki, & da Silva, 1999); (4) *Teoria do Estímulo e Resposta* (Skinner, 1953, 1958, 1959; Keller & Schoenfeld, 1950; Sidman, 1960).

3) *Teoria Psicométrica* ou a *Teoria dos Testes Psicológicos*: Esta teoria trabalha igualmente com dois parâmetros, a saber, a resposta (comportamento) do sujeito e o critério. Pelo fato de que

o critério é entendido de diferentes maneiras, surgem duas teorias psicométricas bastante distintas, quais sejam, a Teoria Clássica dos Testes (TCT), que entende o critério como comportamento (futuro), e a Teoria de Resposta ao Item (TRI), que entende como critério o traço latente (*latent modeling*). Estas serão detalhadas no capítulo 4.

7 – O problema do erro

7.1 – Conceito de erro

A medida é um procedimento empírico e não existe procedimento empírico isento de erro. Esta não é uma afirmação lógica, mas pode ser considerado um postulado e pode ser empiricamente verificada através de operações de mensuração. Mesmo na medida fundamental, é impossível se evitar o erro. Argumentando com Popper (1972), podemos dizer que medir consiste na determinação da coincidência de pontos: um sinal no objeto a ser medido e um sinal no instrumento de medida (metro, por exemplo). Agora, não existe tal coincidência no sentido de que os dois pontos se fundem num ponto único, há apenas uma justaposição dos dois pontos. A precisão perfeita da justaposição só seria finalmente efetuada se pudesse ser verificada num aumento ao infinito desses dois pontos; e acontece que com o aumento deles ao infinito se verifica que os pontos realmente (de fato) não estão perfeitamente alinhados, mas apenas aparecem mais ou menos próximos. Assim, a coincidência se faz dentro de um intervalo: o ponto do corpo medido cai dentro de um intervalo de pontos no instrumento ("extremos de condensação"). Quanto menor este intervalo, maior a precisão da medida. Por esta razão, é costumeiro entre os cientistas apresentar, além do valor da medida, o seu equivalente erro provável, o qual define precisamente estes extremos de condensação.

A esta altura, você provavelmente já sentiu que o número utilizado na medida dos fenômenos naturais não é exatamente o número que os matemáticos estudam, embora ele mantenha importantes características em comum, tais como ordem e até aditividade. É, entretanto, esclarecedor observar que o número estudado pelos matemáticos é um conceito absolutamente claro e distinto; ele é um ponto, ele é o objeto direto de estudo do matemático. Por outro lado, o número utilizado na medida já não é mais um ponto; ele é um intervalo, o que significa que ele pode ser mais ou menos ele mesmo, isto é, ele admite variabilidade, o que é uma manei-

ra elegante de dizer que ele admite erro. Este número "grosseiro" é objeto de estudo de um ramo da Matemática chamada Estatística. Assim, enquanto a Estatística estuda o número como representação de algo diferente dele, porque ele é uma descrição de fenômenos naturais e não mais um conceito original, a Matemática estuda precisamente o número em sua própria identidade.

7.2 – Tipos de erro

Os erros podem ser debitados ou à própria observação ou à amostragem de objetos ou eventos na qual a medida foi realizada.

7.2.1 – Erros de observação

Há quatro fontes principais de erros de observação: (1) erros instrumentais devidos a inadequações do instrumento de observação, (2) erros pessoais devidos às diferentes maneiras de cada pessoa reagir, (3) erros sistemáticos devidos a algum fator sistemático não controlado, como por exemplo medir a temperatura em nível diferente da do mar, e (4) erros aleatórios, que não têm causa conhecida ou cognoscível. Há, inclusive, curiosos acontecimentos neste particular, como a demissão do seu assistente pelo astrônomo real Nevil Maskelyne (Inglaterra) por ter observado a passagem de estrelas e planetas meio segundo depois do que tinha ele mesmo observado. O problema não é tanto a existência desses erros, que são inevitáveis, mas sim identificar as suas fontes e propor meios de reduzi-los. A tabela 2-4 dá uma síntese desta problemática.

7.2.2 – Erros de amostragem

Como a pesquisa empírica normalmente não pode ser feita sobre todos os membros de uma população de eventos ou objetos, tipicamente se seleciona uma amostra destes eventos ou objetos. Esta escolha de indivíduos no meio de uma população é sujeita a desvios, vieses, isto é, erros. O problema não é os erros em si, se o interesse fosse tirar conclusões sobre a amostra selecionada. Acontece, porém, que o interesse do pesquisador é tirar conclusões ou fazer inferências sobre toda a população da qual a amostra foi retirada. Neste caso, o erro de amostragem é desastroso, dado que poderia ocasionar inferências errôneas, considerando a presença de vieses da amostra com respeito a esta população (falta de representativi-

dade). Para solucionar os problemas advindos da seleção da amostra foi desenvolvida a teoria estatística da amostragem.

Tabela 2-4. Erros de medida: fontes e controle

Tipo	Causa	Controle
instrumental	instrumento	calibração
pessoal (observador)	diferenças individuais	atenção, treinamento
sistemático	fator específico	experimental ou estatístico
aleatório	não conhecida	teorias do erro (probabilidade)
amostragem	seleção da amostra	representatividade da amostra (teoria estatística)

7.3 – A teoria do erro

Dado que o erro está sempre presente em qualquer medida e que sua presença constitui uma ameaça séria à tomada de decisões científicas, é de capital importância que haja meios de neutralizar ou diminuir os seus efeitos ou, pelo menos, de conhecer sua grandeza, o mais aproximado possível, para saber o tamanho de risco em que se está incorrendo ao tomar decisões baseadas na medida. Todos os esforços para controlar o erro através de procedimentos experimentais são necessários, mas nem por isso o erro vai desaparecer, dado que a ocorrência dele é imprevisível, isto é, não é nunca possível se determinarem as causas de todos os erros possíveis numa medida. Para enfrentar esta situação foi desenvolvida a teoria do erro, baseada na teoria da probabilidade e dos eventos casualoides.

Um evento casualoide ou aleatório é definido por Popper (1974: 190): "Uma sequência-evento, ou sequência-propriedade, especialmente uma alternativa, se diz 'casualóide' ou 'aleatória' se e somente se os limites das frequências de suas propriedades primárias forem 'absolutamente livres', isto é, indiferentes a qualquer seleção que se apoie nas propriedades de qualquer ênupla de predecessores". Em palavras mais simples, um evento empírico é aleatório se sua ocorrência não pode ser predita a partir dos eventos que ocorreram antes dele, isto é, ele é totalmente independente (livre) com relação ao que aconteceu antes. Imagine o jogo de lançar uma moeda para obter cara ou coroa ou de um dado: qualquer que tenha sido o resultado nos lançamentos anteriores do dado, o resultado (um entre

os seis possíveis) do próximo lançamento é totalmente imprevisível; isto é liberdade absoluta.

O erro na medida é considerado um evento aleatório pela teoria do erro. Feita esta suposição, então é possível tratar o erro dentro da teoria da probabilidade, do teorema de Bernoulli, que baseia a lei dos grandes números e da curva normal, que determina a probabilidade de ocorrência dos vários elementos da série, no nosso caso, da série aleatória composta dos vários tamanhos de erros cometidos na medida.

A curva normal define que uma sequência aleatória de eventos empíricos se distribui normalmente em torno de um ponto modal (média) igual a 0 e uma variância igual a 1. Este valor modal, no caso de uma distribuição de erros, significa que estes se cancelam no final, dado que este valor (0) é o que possui a maior probabilidade na distribuição. Contudo, isto é absolutamente verdadeiro somente na distribuição de uma série aleatória de um número infinito de eventos, segundo o teorema de Bernoulli. Este teorema, na verdade, afirma que um segmento 'x' de elementos de uma série aleatória infinita 'A' (isto é, com liberdade absoluta) que se aproxima da série total ($x \rightarrow A$) possui os mesmos parâmetros desta série. Isto significa que, quanto maior o segmento, mais próximo está dos parâmetros da série ou, em outras palavras, quanto maior o segmento, menor o desvio dos parâmetros dele dos da série. Diz Popper (1974: 198): "Assim, o teorema de Bernoulli assevera que os segmentos mais curtos de sequências casualoides mostram, muitas vezes, grandes flutuações, enquanto que os segmentos longos sempre se comportam de modo que sugerem constância ou convergência; diz o teorema, em suma, que encontramos desordem e aleatoriedade no pequeno, ordem e constância no grande. É a este comportamento que se refere a expressão 'lei dos grandes números'".

Na prática da pesquisa, contudo, o erro da medida é expresso pelo erro padrão da medida (EPM) que é o valor médio da variância, isto é,

$$EPM = \frac{\sqrt{s^2}}{\sqrt{N-1}} = \frac{DP}{\sqrt{N-1}}$$

onde,

s^2 = variância

N = número de sujeitos.

A informação dada pelo erro padrão da medida esclarece que a medida verdadeira de um atributo se situa entre o valor médio das medidas efetuadas e um erro padrão em torno dele (isto é, mais um erro padrão e menos um erro padrão).

8 – Importância da medida

Poder-se-ia perguntar, diante de tantas dificuldades que a medida apresenta, se há vantagem em se utilizar métodos de medições em lugar de métodos puramente qualitativos ou descritivos. Parece que a resposta deva ser positiva, porque aqueles métodos se apresentam superiores a estes em, pelo menos, duas áreas: precisão e simulação.

8.1 – Precisão

Apesar da medida nunca ser destituída de erro, ela é capaz de definir limites dentro dos quais os reais valores dos atributos medidos se encontram. O conceito de pontos de condensação ou de extremos imprecisos (Popper, 1974) nos indica a solução da questão da precisão da medida. Fazer pontos coincidirem (ponto extremo do atributo do objeto a ser medido e ponto de referência do instrumento de medida) significa determinar que o ponto do atributo cai dentro de um intervalo de pontos extremos do instrumento. A questão, então, reduzir-se-ia a determinar estes pontos extremos do intervalo, que, por sua vez, também caem dentro de um intervalo, cujos pontos extremos precisariam ser determinados e, assim, indefinidamente. Isto é, nunca daria para decidir nenhum intervalo de pontos de condensação. Entretanto, os pontos extremos do intervalo de condensação seriam definidos por intervalos cada vez menores, de sorte que se pode finalmente definir um intervalo, o menor possível, com pontos extremos imprecisos, dentro do qual o valor real do atributo se encontra. Assim, fica definido um intervalo mínimo mais provável dentro de seus pontos extremos e, igualmente, a margem de erro tolerada ou provável. De sorte que não se contentaria em simplesmente afirmar que o atributo é mais ou menos de tal magnitude, mas que tem uma magnitude definida dentro de limites (intervalo) assim estabelecidos. A redução ao mínimo do intervalo dos pontos de condensação, evidentemente, depende de avanços tecnológicos no instrumental de medição.

Sendo isso possível, fica mais precisa tanto a descrição do fenômeno natural quanto a comunicação sobre o mesmo. Fica também mais

exata a definição das operações e procedimentos utilizados na observação dos mesmos fenômenos. A medição não torna a observação possível, mas a torna mais unívoca, isto é, menos ambígua, mais precisa. Esta vantagem da medição se torna ainda mais crucial na observação do muito grande (macroscópico) e do muito pequeno (microscópico).

8.2 – A simulação

A manipulação da realidade é geralmente complexa, difícil e custosa. Além disso, às vezes ela é impossível ou eticamente condenável. Por exemplo, não parece aceitável querer estudar os efeitos da bomba atômica sobre uma cidade, explodindo uma. Mas conhecendo com precisão as relações entre os componentes em jogo e suas magnitudes, pode-se utilizar modelos matemáticos para simular os efeitos que queremos estudar e que, de outro modo, seria impossível ou impraticável pesquisar.

Conclusão

A medida em ciências empíricas não pode ser considerada uma panaceia para decidir todos os problemas do conhecimento da realidade, inclusive porque não é ela que define o objeto e nem o método da ciência. Mas, diante das vantagens apresentadas, seria quiçá até irracional não se aproveitar da medida como instrumental de trabalho no estudo da realidade. A história da ciência parece demonstrar, inclusive, que o avanço do conhecimento científico está ligado ao maior ou menor uso da medida, sobretudo quando ela está baseada numa teoria axiomatizada, isto é, quando há a explicitação clara do maior número possível dos axiomas necessários. Infelizmente, na medida em ciências psicossociais, esta axiomatização está longe de ser uma realidade. Mesmo assim, a discussão sobre a viabilidade da medida nestas ciências parece uma disputa mais inócua que produtiva; uma discussão de como se proceder à medida parece mais substantiva, produtiva e útil para o desenvolvimento destas ciências.

CAPÍTULO 3
A medida psicométrica

1 – Introdução

A problemática da medida em Psicologia ainda suscita questionamentos ardorosos entre os psicólogos. Há três décadas, o próprio Guttman (1971) ainda se interrogava o que exatamente significava "medida" em ciências psicossociais. Embora, nestas ciências, fossem correntes as expressões sociometria, antropometria, biometria, psicometria, econometria e outras 'metrias', continuavam dúvidas sobre sua significação no campo da epistemologia e da metodologia. Os vários prefixos das "-metrias" evidentemente revelavam a área de conteúdo em que a medida era aplicada. Assim, psicometria seria o uso da medida em Psicologia.

Ainda hoje, essa situação levantada por Guttman não está de todo resolvida. De fato, a teoria da medida em ciências não constitui campo pacífico entre os pesquisadores, sobretudo em ciências psicossociais. Outro complicador, neste contexto, é a tendência de alguns em reduzir, por exemplo, a psicometria, cuja preocupação central é a construção e verificação de hipóteses científicas, à psicoestatística, cuja preocupação é a inferência a partir de amostras. Aliás, este tipo de divergência foi o que provocou, em análise fatorial, a divisão do grupo de Thurstone dos anos de 1930 em várias correntes, cada qual seguindo seus interesses pessoais de psicometristas, de estatísticos ou de matemáticos, inclusive com a criação de revistas especializadas divergentes da *Psychometrika*.

Este capítulo pretende caracterizar a Psicometria dentro de uma orientação epistemológica quantitativista, mas como ramo das ciências empíricas e não das matemáticas. Estas duas não são conflitantes, mas são epistemologicamente independentes. A distinção precisa ser defendida e cobrado o ônus da prova para a justificativa da viabilidade da associação das duas, isto é, ciência, de um lado, e matemática, do outro (cf. cap. 2).

Em seu sentido etimológico, Psicometria seria, conforme insinuou Guttman (1971), toda a classe de medida em Psicologia, similarmente a sociometria ser na sociologia, econometria na economia, etc. Em seu sentido mais restrito, e é neste que ela é normalmente entendida, Psicometria constitui uma das várias formas de medição em Psicologia. Ela é uma das formas de medida por teoria (cf. cap. 2), onde se situam igualmente a teoria dos jogos e a teoria da detecção do sinal.

A teoria que fundamenta a Psicometria neste sentido estrito assume os postulados da teoria da medida em geral. Ademais, como foi desenvolvida sobretudo por estatísticos, ela usa símbolos que expressam parâmetros, os quais representam variáveis de caráter abstrato, o que é suficiente para desenvolver o modelo matemático da teoria. Contudo, como a Psicometria é um ramo da Psicologia e não da Estatística, tais parâmetros precisam adquirir uma definição substantiva em termos da disciplina psicológica, não sendo suficiente sua intelecção em termos puramente estatísticos. Assim, quando se falar, por exemplo, de variáveis hipotéticas, tais como teta ou traço latente, estas expressões devem assumir conteúdo psicológico, porque a Psicologia não tem, como objeto de estudo, parâmetros e sim processos comportamentais, processos psíquicos... No caso da Teoria Clássica dos Testes (TCT), os parâmetros envolvidos são comportamentos; no caso da Teoria de Resposta ao Item (TRI), além dos comportamentos entram processos definidos como traços latentes. Este último conceito não é imediatamente óbvio e, por isso, se fazem necessários alguns esclarecimentos sobre o mesmo. Há aqui, na verdade, um fundo de caráter epistemológico que opõe a TRI a TCT, oposição que se fulcra numa concepção mais básica do próprio ser humano e que precisa ser explicitamente esclarecida, como será feito a seguir, juntamente com outros conceitos relevantes dentro da Psicometria, quais sejam os de sistema, propriedade, magnitude, bem como a representação comportamental da estrutura latente (*latent trait modeling*).

2 – Comportamento vs. traço latente

A TCT surgiu dentro da concepção monista materialista que imperava nas ciências em geral desde o empiricismo inglês do século XVII, enquanto a TRI faz suposição de ou, pelo menos, permite se fundamentar uma concepção dualista interacionista do ser humano. A figura 3-1 procura ilustrar esta problemática.

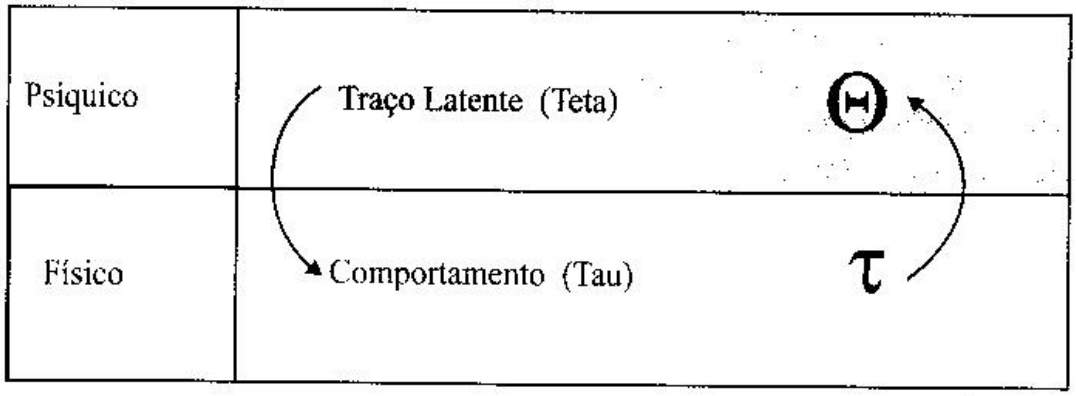


Figura 3-1. Ilustração da concepção dualista do ser humano

A Psicometria tradicional (TCT) vai definir a qualidade dos testes psicológicos (que são estímulos comportamentais ou, se você quiser, variáveis observáveis) em termos de um critério, sendo este representado também por comportamentos, a saber, comportamentos presentes ou futuros. Desta forma, a TCT trabalha exclusivamente com comportamentos: tanto o teste (a medida) quanto o critério são comportamentos, realidades físicas. Ela trabalha unicamente no nível do tau (τ) do ser humano. Por outro lado, a TRI define a qualidade dos testes (que são comportamentos ou variáveis observáveis) em função de um critério que não é o comportamento e sim variáveis hipotéticas, as quais chama de teta ou traço latente (θ). Formalizando estes enfoques, você pode dizer que na

TCT: comportamento (teste) = f (critério), isto é, as tarefas do teste se definem em função de outros comportamentos (presentes ou futuros: o critério) que o teste pretende predizer;

TRI: comportamento = f (θ), isto é, as tarefas do teste se definem pela, são efeito da, aptidão ou traço latente (o θ).

Entretanto, como certas características do próprio item ou tarefa do teste, bem como os erros de medida afetam a qualidade de um teste, as equações acima na verdade se apresentariam conceitualmente da seguinte forma:

TCT: teste = f (critério, item, erro)

TRI: teste = f (teta, item, erro)

Onde,

critério = comportamentos presentes ou futuros

teta = aptidão, traço latente

item = características do item (dificuldade, discriminação, ...)

erro = erros de medida.

Uma nota: alguns psicólogos, particularmente os de orientação behaviorista, acenam ao fato de que a distinção entre processo mental (teta) e comportamento (tau) é inócua, porque o processo mental também é um comportamento. Isto me parece um puro jogo de palavras, isto é, igualar processo mental e comportamento. O problema não está nas palavras, mas nos fundamentos epistemológicos que tais expressões (teta e tau) querem insinuar e nos quais estão alicerçados. Isto quer dizer o seguinte: O psicólogo (note que digo psicólogo e não psicometrista, para evitar que se confunda psicometria como um suposto ramo da estatística em vez da psicologia) que trabalha com teta e tau quer defender, por necessidade, uma visão dualista do ser humano (mente e corpo), enquanto que aquele que trabalha somente com o tau é, por necessidade, monista, e, no caso, monista materialista (cf. a discussão sobre traço latente, mais adiante).

3 – Traço latente

O conceito de traço latente não é isento de ambiguidades e controvérsias entre os autores que trabalham com tal construto. A variedade de expressões utilizadas para representá-lo já indica tal dificuldade. Traço latente vem referido ou inferido sob expressões como: variável hipotética, variável fonte, fator, construto, conceito, estrutura psíquica, traço cognitivo, processo cognitivo, processo mental, estrutura mental, habilidade, aptidão, traço de personalidade, processo elementar de informação, componente cognitivo, tendência, atitude e outros. Alguns autores (Weiss, 1983) incluem até o escore verdadeiro da Teoria Clássica dos Testes como sendo um traço latente[1]. Para o estatístico, isso pode fazer sentido, já que eles trabalham somente com parâmetros; neste caso, o escore bruto do sujeito (resultado num teste) seria função de um escore abstrato, chamado precisamente de escore verdadeiro (que sendo abstrato seria uma variável não observável e, portanto, um traço latente). Mas nós queremos falar como psicólogos e não como estatísticos; como consequência, esse modo de en-

1. Sobre escore verdadeiro, cf. capítulo 4.

carar o traço latente não resolve o problema para o psicólogo, cujo objeto de estudo são processos psicológicos e não parâmetros abstratos. Assim, vamos encarar traço latente como processo psicológico. Isto evidentemente não resolve os problemas; aliás, cria dificuldades maiores ainda, pois já a própria *natureza ontológica* de traço latente deixa dúvidas: deve ele ser concebido como um rótulo, representando uma síntese hipotética de um conjunto de comportamentos reais, ou como uma realidade mental? Para este autor, o conceito faz mais sentido quando entendido como realidade na concepção popperiana de que é real aquilo que age sobre coisas consideradas reais, como as coisas físicas materiais: "Deve-se então admitir que as entidades reais podem ser concretas ou abstratas em vários graus. Em física, aceitamos forças e campos de força como reais, pois agem sobre coisas materiais. Mas essas entidades são mais abstratas e, talvez, também mais conjeturais ou hipotéticas do que são as coisas materiais comuns. Forças e campos de força são ligados a coisas materiais, a átomos e a partículas. Têm um caráter dispositivo: são tendências para interagir. Podem assim ser descritas como entidades teóricas altamente abstratas, nós as aceitamos como reais, quer elas ajam de forma direta ou indireta sobre as coisas materiais" (Popper & Eccles, 1977: 27-28). Esta posição não pode ser considerada uma concepção platonista, no sentido de que o traço latente seria a coisa realmente real e o resto (comportamento, por exemplo) a sombra. Considerar a realidade do traço latente como platonismo é visão do behaviorismo radical, para o qual somente é real o que os sentidos são capazes de verificar, postulado não necessário para uma concepção empiricista de ciência. Empírico não se confunde com este behaviorismo radical. Mesmo defendendo o traço latente como realidade, com o mesmo direito que o comportamento físico o é, nem por isso se defende que o estudo científico, isto é, empírico, do traço latente deva ser feito diferentemente do estudo científico que se faz do comportamento. A solução deste aparente enigma está em que o traço latente, para ser cientificamente estudado, deve ser representado em comportamentos. Como assim? Vejamos.

As estruturas latentes são atributos impérvios à observação empírica, que é o método da ciência. Então, o traço latente precisa ser expresso em comportamentos para ser cientificamente abordado. Porque sendo o comportamento (verbal, motor) o único nível em que se pode trabalhar cientificamente (empiricamente) em Psicologia, é neste nível que se deve procurar a solução para o problema da representação e, portanto, do co-

nhecimento dos processos latentes. A teoria que fundamenta o isomorfis-mo comportamento – processos latentes é o fulcro epistemológico da Psicometria, juntamente com a concepção de processos latentes como dimensões, isto é, atributos mensuráveis. Postula-se que, ao se operar sobre o sistema comportamento, está-se operando sobre os traços latentes (isomorficamente). Assim, a medida que se faz ao nível comportamental é a medida dos traços latentes. Como o comportamento representa estes traços latentes? É o problema das definições operacionais. A Psicometria responde a esta questão pela análise de uma série de parâmetros que os comportamentos (tipicamente chamados de itens) devem apresentar. O parâmetro fundamental da medida psicométrica (escalas, testes, ...) é a demonstração da adequação da representação, isto é, a demonstração do isomorfismo entre a ordenação nos procedimentos empíricos e a ordenação nos procedimentos teóricos do traço latente (é a problemática da *validade*). Significa demonstrar que a operacionalização do atributo latente em comportamentos (itens) de fato corresponde a este atributo. Esta demonstração é tipicamente tentada através de análises estatísticas dos itens individualmente e do teste em seu todo. Para tanto, a comunidade científica desenvolveu uma série de parâmetros mínimos que a medida psicométrica deve apresentar para se constituir em instrumento legítimo e válido, como veremos nos capítulos subsequentes deste livro.

Além disso, traço latente pode ser entendido como um simples parâmetro de caráter matemático ou estatístico, como parece entendê-lo a Psicometria Moderna da Teoria da Resposta ao Item. Para fins de trabalhar com o modelo psicométrico, tal modo de conceber traço latente é suficiente. Mas não parece suficiente para o cientista psicólogo, porque a este interessa o estudo de processos psicológicos e não de puros conceitos e parâmetros matemáticos. Consequentemente, é preciso se dar ao construto de traço latente algum *conteúdo psicológico* ou, então, cair na solução do behaviorismo radical e entender este construto como um puro rótulo.

Além desta controvérsia, existem diferentes maneiras de conceber traço latente quando se trata de definir sua *estrutura elementar*. Na verdade, há aqui duas tendências distintas e em vários níveis: concepção elementarista (reducionista) e concepção estruturalista, conforme detalhado na figura 3-2.

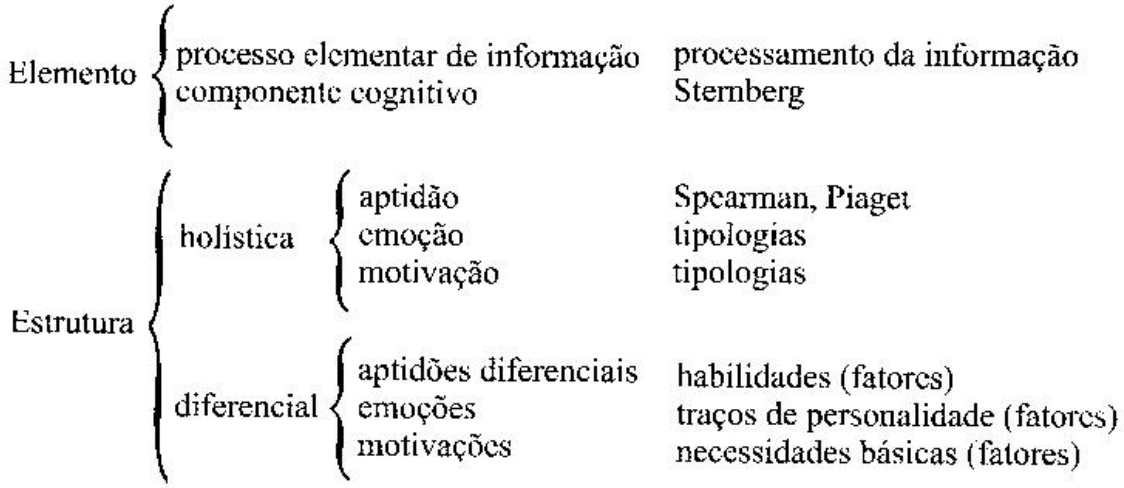


Figura 3-2. Visões elementarista e estrutural de traço latente

A figura 3-2 permite visualizar os teóricos que entendem traço latente como uma estrutura global, seja constituindo toda a psique do ser humano ou grandes conjuntos dela. No caso das aptidões humanas, por exemplo, Spearman defende a teoria do fator intelectual único (fator G – Spearman, 1904b); Piaget (1952) fala do desenvolvimento das estruturas cognitivas. Na área da emoção e da motivação aparecem as tipologias de tipo Jung (1921), Kretschmer (1925) e Sheldon (1940, 1942), bem como as teorias modernas do temperamento. Estas concepções apresentam uma tendência de considerar os traços latentes como grandes estruturas que variam de sujeito para sujeito. Dentro ainda de uma concepção estrutural, outros autores concebem os traços latentes de uma forma mais diferenciada, quando falam de fatores. É a tradição na orientação da análise fatorial em Psicometria, onde os fatores são concebidos como variáveis-fonte responsáveis pela qualidade da execução das tarefas comportamentais. Embora pareçam já elementares, os fatores apresentam ainda um caráter globalizante, dado que não expressam processos cognitivos elementares, mas sim um possível conjunto destes que são necessários para a execução de uma tarefa concreta. Pelo menos, esta é a crítica que Sternberg (1977) faz desta concepção fatorista do traço latente. Sternberg, na verdade, concebe traço latente como algo elementar, isto é, o último elemento cognitivo a que se pode reduzir uma atividade cognitiva, os processos cognitivos (o autor trabalha na área das aptidões). A teoria do processamento da informação (Newell & Simon, 1972) leva ainda mais longe este elementarismo, defendendo o conceito de processo elementar de informação como sendo

o processo mais elementar possível no processamento da informação, o qual não pode ser analisado em elementos menores.

Para ilustrar, brevemente, estas várias concepções, no caso dos processos cognitivos ou das habilidades e aptidões, podemos considerar a tabela 3-1.

Tabela 3-1. Enfoques conceituais de processo cognitivo

Enfoque	Traço latente	Analogia ilustrativa	Referência
Processamento da informação	processo elementar de informação – eip	elemento atômico da Física Nuclear	Newell & Simon, 1972
Psicologia Cognitiva	componente cognitivo	elemento da tabela periódica da Química	Sternberg, 1977
Psicometria	fator	elemento natural (Geologia, Geografia, Anatomia, etc.)	Fatoristas (Thurstone, Cattell, Guilford,...)

Assim, as concepções de traço latente dependem do nível de especificidade que se quer dar a este construto ou parâmetro. Os fatoristas estão mais interessados em chamar de traço latente aquele conjunto de processos cognitivos necessários para a execução de uma tarefa (de fato, um barramento correlacionado de processos ou módulos), falando de habilidades primárias, que seriam combinações de processos cognitivos elementares, isto é, de representações mentais de objetos e símbolos. O fator seria um sistema de processos cognitivos ou de componentes cognitivos. Ao contrário, Sternberg chama de processo cognitivo estas mesmas representações mentais individuais, que serão os componentes cognitivos. Agora, para representar mentalmente objetos e símbolos, uma série de processos ainda mais elementares é necessária e, então, estes sim seriam finalmente os processos elementares básicos do processamento da informação para Newell e Simon, elementos que se combinam num sistema de processos da informação para a explicação de uma tarefa comportamental. Onde parar nesta tendência reducionista? Sternberg (1977: 65-66) afirma: "o componente não é necessariamente, e normalmente não o é, a unidade mais elementar de comportamento que se possa estudar. Operações que são consideradas sem importância dentro da teoria são especificadas no

modelo do processamento da informação do desempenho de uma tarefa, mas não serão identificadas como componentes separados. A razão para esta seletividade é que tarefas complexas podem requerer centenas ou até milhares de operações, a maioria das quais se apresentam desinteressantes do ponto de vista da teoria".

Argumentação similar a esta de Sternberg pode ser feita pelo fatorista. É interessante para o fatorista parar, neste reducionismo, no nível do fator. Este é concebido como um "macro" ou uma rotina de execução de tarefas. Qual a vantagem desta concepção? A vantagem em conhecer este nível de operar dos sujeitos consiste em que, na sua grande maioria (a média), os sujeitos se comportam em termos de habilidades desenvolvidas (macros) em sua vida; ao se depararem com as situações comuns da vida, os sujeitos lançam mão de seus esquemas de resposta desenvolvidos (as habilidades, os macros, as receitas) sem ter que, a cada nova situação, desenvolver novas técnicas e táticas de ação. Assim, é importante para talvez a maioria das situações da maioria dos sujeitos conhecer estas estratégias (habilidades, macros, rotinas) e assim predizer suas chances de êxito na solução dos problemas comuns da vida. É o que o fatorista ou o psicometrista é capaz de realizar. "A coisa fica preta" quando estes mecanismos, estas estratégias não mais funcionam ou não funcionam direito, como ocorre nos casos que a Psicologia dos deficientes ou dos superdotados estuda. Neste caso, o que precisa ser revisto é a própria estratégia ou macro do sujeito. Para tanto é necessário desmontar (estudar) o próprio macro e descobrir qual processo nele não está funcionando corretamente e remontá-lo corrigido. Mas esta tarefa exige trabalhar com componentes cognitivos do tipo Sternberg. O puro treinamento da estratégia (do fator, da aptidão), o que seria a recomendação de um fatorista, nem sempre levaria a resultados satisfatórios. O que implica em dizer que a visão fatorista de traço latente funciona na grande maioria dos sujeitos e na grande maioria das situações, mas pode falhar em situações específicas, nas quais uma concepção mais cognitivista seria mais eficaz, mas esta já ou ainda não é a visão psicometrista.

Além da diferença no nível de reducionismo, outra vertente importante de diferenças entre estes vários sistemas de conceber o traço latente consiste na visão mais *estruturalista* das concepções holísticas, que tendem a considerar os traços latentes como entidades. As concepções mais elementaristas tendem a considerar traço latente como *processos*. Assim, Newell e Simon consideram como processos elementares de in-

formação (*eips*) a discriminação, a testagem e a comparação, a criação de símbolos, a criação de estruturas de símbolos, produção de respostas externas em função de estruturas simbólicas internas, designação de estruturas simbólicas, e memorização de estruturas simbólicas. Por sua vez, Sternberg fala de processos de codificar, inferir, mapear, aplicar, justificar e responder. As diferenças individuais que ocorrem nestes processos seriam devidas à dificuldade e duração que diferentes sujeitos encontram ou necessitam para eliciar estes processos, enquanto para os fatoristas, por exemplo, as diferenças surgiriam principalmente em função da magnitude (tamanho, dimensão, quantidade) do traço latente possuído por diferentes sujeitos. A Psicometria trabalha com o conceito fatorista de traço latente.

4 – Sistema

O sistema representa o objeto de interesse, chamado também de objeto psicológico. A Psicometria moderna enfoca como seu objeto específico as estruturas latentes, os traços psicológicos. Ela teoriza a partir destas estruturas hipotéticas. Deste enfoque, evidentemente, surgem dificuldades, dado que a ciência empírica, dentro da qual a Psicologia se define, tem como objeto de conhecimento os fenômenos naturais abordados através da observação, que, no caso da Psicologia, é o comportamento. Este problema será abordado na seção da representação comportamental da estrutura latente. Aqui é relevante salientar que a Psicometria trabalha com a teoria dos traços latentes, sendo, portanto, as estruturas psicológicas latentes o 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 em sistema como o objeto imediato de interesse dentro de um delineamento de estudo e não é uma entidade ontológica monolítica e unívoca.

5 – Propriedade

Um sistema apresenta atributos que são os vários aspectos ou propriedades 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 ou atributos que definem os mesmos, sendo estes atributos o foco imediato de observação ou 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.

6 – Magnitude

A Psicometria assume, ainda, que estes atributos psicológicos apresentam magnitude: os atributos são dimensões, isto é, são mensuráveis. Trata-se do conceito de quantidade: os atributos ocorrem com quantidades definidas e diferentes de indivíduo para indivíduo. Quantidade é um conceito matemático que se define em função dos axiomas de ordem e de aditividade dos números: os números não somente são diferentes, mas um é maior que outro, de sorte que eles podem ser ordenados numa série monotônica crescente de magnitude. Ao se falar de magnitude dos atributos empíricos, quer se referir, pelo menos, a esta propriedade numérica de ordem crescente. Digo pelo menos, porque nem sempre é possível salvar na medida os axiomas da aditividade que implicam na possibilidade de concatenação, resultando em medida de nível intervalar ou de razão. Aliás, é esta suposição de magnitude das propriedades psicológicas que torna interessante a utilização do modelo matemático no estudo dos fenômenos de que trata a Psicologia.

7 – O problema da representação comportamental

Mesmo se admitindo que as estruturas latentes tenham atributos e que estes possuam magnitude, fica o problema fundamental de que estes atributos são impérvios à observação empírica que é o método da ciência. Então como fica a utilidade de todo este teorizar? Estamos aqui nos defrontando com o problema da representação: qual é a maneira adequada de se representarem 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 (verbal, motor) é o único nível em que se pode trabalhar cientificamente (empiricamente) em Psicologia, é neste nível que se deve procurar a solução para o problema da representação e, portanto, do conhecimento dos processos latentes. Está ali também o problema básico da Psicometria: a legitimidade de suas operações depende da legitimidade desta representação. A teoria que fundamenta o isomorfismo comportamento – processos latentes é o fulcro epistemológico da Psicometria, juntamente com a concepção de processos latentes como dimensões, isto é, atributos mensuráveis. Postula-se que, ao se operar sobre o sistema comportamento, está-se operando sobre os traços latentes (isomorficamente). Assim, a medida que se faz ao nível comportamental é a medida dos traços latentes.

Como o comportamento representa estes traços latentes? É o problema das definições operacionais. A Psicometria responde a esta questão pela análise de uma série de parâmetros que os comportamentos (tipicamente chamados de itens) devem apresentar individualmente e em grupo (um teste), como veremos a seguir.

7.1 – Os parâmetros individuais dos itens

Na avaliação de cada item são verificadas as seguintes características (que serão detalhadas no cap. 5):

1) *Modalidade*: em termos de seu conteúdo, os comportamentos (itens) podem ser de tipo verbal ou motor. Dentro destes, pode-se distinguir outros. No caso do verbal, por exemplo, o item pode ser verbal propriamente, numérico, espacial, abstrato, etc., dependendo do conteúdo semântico sobre o qual o comportamento opera poder ser de palavras, números, dimensões espaciais, etc. Pode ser também mais ou menos abstrato, dependendo do nível de universalidade dos conceitos envolvidos: conceitos singulares, universais de menor abstração, universais de maior abstração. Neste particular, a Psicometria deveria interagir com a Psicolinguística, já que interfaceia com o campo do significado.

2) *Saturação*: o comportamento humano tipicamente se apresenta como multimotivado, dado que fatores múltiplos entram na sua aparição, sendo, portanto, difícil, se não impossível, determinar causas ou fatores únicos para qualquer comportamento, ao menos de adultos. Isto implica que seria impossível definir comportamentos (itens) críticos para qualquer traço latente, no sentido de um comportamento 'x' ser específico e único

de tal traço e não interfaceando com qualquer outro traço. Podemos dizer, então, que somente parte do comportamento 'x' representa o traço, ele covaria com o traço; mas esta covariação não constitui toda a variância do 'x'. É, por isso, importante descobrir o nível desta covariância 'x'-e-traço latente em questão. Tipicamente tal covariância se expressa estatisticamente através da sua carga ou saturação fatorial, que pode variar de zero a um (positivo ou negativo), sendo que, no caso de ser zero, o comportamento seria uma representação equivocada, inadequada, do traço. Este parâmetro se relaciona a unidimensionalidade dos instrumentos psicológicos de medida (cf. cap. 11).

3) *Dificuldade* (complexidade): um comportamento é mais difícil ou mais complexo na medida em que ele exige maior nível de magnitude do traço em questão para ser eficaz ou corretamente executado. A expressão "dificuldade" se originou dentro da medida das aptidões e é mantida, por exemplo, no parâmetro *b* da Teoria de Resposta ao Item (TRI), mesmo quando se trata da medida de atitudes ou traços de personalidade em geral. Talvez a expressão "complexidade" fosse mais adequada para representar este parâmetro, na medida em que ela especifica que um comportamento é mais complexo e, portanto, mais difícil, porque a sua correta execução (no caso de se tratar de aptidão cognitiva) ou a adesão a seu conteúdo semântico (no caso de traços de personalidade e atitudes) depende de um maior nível de magnitude no traço latente. O que exatamente torna um item mais complexo é ainda tema de pesquisa, do qual a Psicologia Cognitiva vem se interessando bastante como forma de estudar os processos cognitivos. A Psicometria avalia este parâmetro através de técnicas puramente estatísticas, mas seria de enorme valor a descoberta dos elementos que constituem maior complexidade no item, sobretudo para fins de construção do próprio elenco de itens da medida dos traços latentes. Este parâmetro afeta a amplitude de uma escala de medida: o elenco de itens cobre adequadamente toda a extensão de magnitudes possíveis de um dado traço ou somente um segmento delas e qual segmento?

4) *Discriminação*: o poder discriminativo de um item (comportamento) se define como a capacidade que ele apresenta de separar (discriminar) sujeitos com magnitudes próximas do mesmo traço (teta). Quanto mais extremas devam ser as magnitudes do atributo para que o item possa discriminá-las, menos discriminativo ele é e vice-versa. A TRI define como *a* este parâmetro. Que característica do item determinaria seu poder discriminativo? Novamente a Psicologia Cognitiva poderia lançar luzes

nesta questão, definindo os elementos cognitivos que a reação a um item utiliza. Seria a univocidade semântica do item, isto é, um sentido bem definido com nível reduzido de ruído, a saber, conceitos despojados de conotações? Uma informação desta natureza auxiliaria grandemente a construção de itens comportamentais mais típicos e adequados para a medida dos traços.

5) *Viés de resposta*: mesmo os itens apresentando bons índices nos parâmetros acima descritos, há toda uma série de dificuldades que aparecem afetando a qualidade da resposta do sujeito aos itens, dificuldades estas que provêm de fatores subjetivos do respondente e que poderiam ser agrupadas dentro do conceito de tendências. Tendência seria uma atitude, consciente ou não, do sujeito responder de maneiras sistemáticas alheias ao conteúdo semântico dos itens. São os erros de resposta devido ao responder ao acaso, dar respostas estereotipadas (sempre nos extremos de uma escala ou no ponto neutro), dar respostas em função de supostas expectativas dos outros (desejabilidade social) ou em função de uma ideia preconcebida sobre o objeto de avaliação (efeito de halo), etc. Vários destes problemas podem ser parcialmente evitados, caso seja possível desvendar os fatores sistemáticos responsáveis pelas respostas estereotipadas. Assim, a TRI é capaz de contornar o problema das respostas dadas ao acaso (parâmetro *c*); o formato das escalas de resposta pode reduzir a ocorrência de erros do tipo respostas extremadas ou neutras etc.

7.2 – *Parâmetros do teste (grupo de itens)*

O parâmetro fundamental da medida psicométrica (escalas, testes, ...) é a demonstração da adequação da representação, isto é, a demonstração do isomorfismo entre a ordenação nos procedimentos empíricos e a ordenação nos procedimentos teóricos do traço latente. Significa demonstrar que a operacionalização do atributo latente em comportamentos (itens) de fato corresponde a este atributo. Esta demonstração é tipicamente tentada através de análises estatísticas dos itens individualmente e da escala em seu todo. Infelizmente a literatura neste particular não mostra muita preocupação com a formulação de uma teoria clara, muito menos, axiomatizada, sobre o atributo, a qual pudesse permitir uma elaboração mais bem delineada e planejada de uma escala de comportamentos pertinentes ao atributo. Possivelmente esta situação se deva: (1) à predominância de um enfoque ateórico baseado quase exclusivamente na análise de um elenco de itens, coletado mais ou menos ao acaso ou intuitivamente, em vez de uma pesquisa dos

elementos cognitivos envolvidos nos processos do atributo psicológico e também (2) ao fato de que o desenvolvimento da Psicometria tem sido preponderantemente viabilizado por pesquisadores, cuja formação e preocupações eram mais de estatísticos do que de psicólogos. O desenvolvimento da pesquisa da Psicologia Cognitiva, particularmente do tipo Sternberg (1977, 1979, 1980) e das pesquisas feitas no centro de Pittsburgh (Mulholland, Pellegrino, & Glaser, 1980; Pellegrino, Mumaw & Shute, 1985; Carpenter, Just, & Shell, 1990), deverão auxiliar substancialmente para remediar ou resolver este problema. Os trabalhos de Guilford (1959) também devem ser mencionados neste particular. No momento, em Psicometria, se insiste ainda de maneira exclusiva numa solução estatística. Por outro lado, as dicas que a Psicologia Cognitiva tem, no momento, a dar nesta área da instrumentação psicométrica são ainda muito precárias ou, pelo menos, muito pouco sistematizadas, para servir de base na elaboração e análise dos instrumentos psicológicos.

De qualquer forma, a comunidade científica desenvolveu uma série de parâmetros mínimos que a medida psicométrica deve apresentar para se constituir em instrumento legítimo e válido. Os parâmetros mais básicos se referem, além da análise dos itens (dificuldade e discriminação), à validade e à confiabilidade do instrumento, que constituem temas centrais da Psicometria e que serão detalhados nos caps. 6 e 7.

CAPÍTULO 4
Os modelos da Psicometria: TCT e TRI

Introdução

A Psicometria procura explicar o sentido que têm as respostas dadas pelos sujeitos a uma série de tarefas, tipicamente chamadas de itens. A Teoria Clássica dos Testes (TCT) se preocupa em explicar o resultado final total, isto é, a soma das respostas dadas a uma série de itens, expressa no chamado escore total (T). Por exemplo, o T em um teste de 30 itens de aptidão seria a soma dos itens corretamente acertados. Se você dá 1 para um item acertado e 0 para um errado, e o sujeito acertou 20 itens e errou 10, seu escore T seria de 20. A TCT, então, se pergunta o que significa este 20 para o sujeito? A TRI, por outro lado, não está interessada no escore total em um teste; ela se interessa especificamente por cada um dos 30 itens e quer saber qual é a probabilidade e quais são os fatores que afetam esta probabilidade de cada item individualmente ser acertado ou errado (em testes de aptidão) ou de ser aceito ou rejeitado (em testes de preferência: personalidade, interesses, atitudes). Desta forma, você vê que a TCT tem interesse em produzir *testes* de qualidade, enquanto a TRI se interessa por produzir *tarefas* (itens) de qualidade. No final, então, você tem ou testes válidos (TCT) ou itens válidos (TRI), itens com os quais você poderá construir tantos testes válidos quantos quiser ou o número de itens permitir. Você logo percebe, também, que a riqueza na avaliação psicológica, dentro do enfoque da TRI, consiste em se conseguir construir armazéns de itens válidos para avaliar os traços latentes, armazéns estes chamados de bancos de itens (sobre isto falaremos mais adiante). A seguir serão detalhados estes dois modelos da Psicometria.

I – O MODELO DA PSICOMETRIA CLÁSSICA: TCT

1 – Introdução

Antes de detalhar o tema da Teoria Clássica dos Testes (TCT), é preciso observar que este modelo se ressente de alguns vieses que, embo-

ra não eliminando a força explanatória do mesmo, impõem considerações sobre possíveis limitações do modelo. Estes vieses estão na origem histórica da própria Psicometria, a saber: (1) a Psicometria foi desenvolvida em cima de dados obtidos exclusivamente da medida das aptidões humanas (inteligência); (2) foi elaborada por psicólogos de preocupação predominantemente estatística e (3) sua orientação se insere dentro da visão materialista[1] que dominava as ciências na época. O primeiro viés impõe algumas restrições quanto à aplicação do modelo à medida de outros traços latentes do ser humano que não sejam as aptidões, onde não existem respostas (comportamentos) certas e erradas. O segundo e o terceiro viés resultaram em que a Psicometria, em parte, perdeu de vista a teoria psicológica, o que vai ter impacto importante sobretudo na concepção do parâmetro de validade dos testes, embora ela se adapte perfeitamente ao conceito de precisão da medida, como veremos.

De qualquer forma, vamos inicialmente entender a Psicometria dentro das restrições acima apontadas, focalizando a tarefa que se propôs. A situação concreta com a qual os teóricos da Psicometria se enfrentavam era a seguinte: um sujeito responde a uma série de itens (tarefas) e recebe um ponto por cada tarefa corretamente respondida, obtendo, no final, um *escore total* que é a soma destes pontos ou respostas corretas. O que representa este escore do sujeito? Supostamente ele está expressando uma magnitude que seria a magnitude daquilo que o teste queria medir no sujeito. Contudo, toda e qualquer operação empírica, se sabe, é sujeita a erros. Consequentemente, este escore bruto do sujeito não pode ser a expressão pura da magnitude daquilo que o teste queria medir no sujeito, mas deve conter igualmente uma porção de erros.

Observe que o foco de interesse da TCT não é o traço latente e sim o comportamento ou, melhor, o escore num teste, sendo este teste um conjunto de comportamentos. Diríamos que o enfoque está no tau (τ) e não no teta (θ), entendendo como tau o escore num teste e o teta como o traço latente. O uso da palavra "aptidão" no contexto da TCT (correntemente utilizada) parece indevido, porque sugere o conceito de traço latente, com o qual esta teoria não trabalha. De fato, aptidão na TCT está

1. Materialismo, em sua forma extrema, consiste na crença ou posição filosófica de que toda a realidade se reduz à matéria, isto é, ela se reduz a propriedades físicas.

substituindo a expressão de que o teste está medindo "aquilo que supostamente deve medir"; e este "aquilo que", na TCT, é o critério, como definido no capítulo 3. De sorte que você deve entender aptidão sempre neste sentido, dentro da TCT, ou seja, a capacidade preditiva (preditividade) do teste em referência ao critério, sendo este representado por comportamentos outros ou futuros, e não a representatividade (física) que o teste seria de uma realidade latente[2].

O apanhado histórico elaborado no capítulo 1 mostrou que a Psicometria dispunha de um campo rico de dados e eventos onde se desenvolver e responder a tais perguntas. Foi, entretanto, Spearman (1904, 1907, 1913) que iniciou o desenvolvimento do modelo teórico da Psicometria que, por se apresentar aparentemente simples e linear, favoreceu a aceitação e o uso da jovem Psicometria.

2 – O modelo[3]

O modelo desenvolvido por Spearman implica em alguns postulados básicos, que podem, inclusive, ser considerados as regras do jogo ou as definições iniciais da teoria. Seguindo a clássica síntese de Gulliksen (1950), é preciso, primeiramente, distinguir três componentes neste jogo, a saber, o escore bruto ou empírico (T), o escore verdadeiro (V) e o erro (E). Além disso, é necessário fazer algumas suposições entre as relações existentes entre estes três componentes. Então temos

T = escore bruto ou empírico do sujeito, que é a soma dos pontos obtidos no teste

V = escore verdadeiro, que seria a magnitude real daquilo que o teste quer medir no sujeito e que seria o próprio T se não houvesse o erro de medida

E = o erro cometido nesta medida.

---

2. Note que na TRI o "aquilo que" será o θ e, assim, aptidão no teste significa a representatividade do item em relação ao teta e não a preditividade dele em relação ao critério.

3. No Apêndice A são apresentadas algumas dicas de como proceder na dedução de fórmulas matemáticas usadas neste e outros capítulos.

O primeiro postulado, e que constitui o modelo fundamental da Psicometria Clássica, é de que o escore bruto do sujeito é a soma do escore verdadeiro e do erro, ou seja,

$$T = V + E \qquad (4.1)$$

O escore empírico é a soma do escore verdadeiro e do erro

e, consequentemente, E = T – V, bem como, V = T – E.

A figura 4-1 mostra a relação entre estes vários elementos do escore empírico, onde se vê que este é a união do escore verdadeiro (V) e do erro (E), ou seja, o escore empírico ou bruto do sujeito (T – resultado no teste, conhecido como o escore tau – τ) é constituído de dois componentes: o escore real ou verdadeiro (V) do sujeito naquilo que o teste pretende medir e o erro (E) de medida, este sempre presente em qualquer operação empírica. Em outras palavras, estamos aqui assumindo que, diante do fato de que o escore bruto do sujeito difere do seu escore verdadeiro, esta diferença é devida ao erro ou, melhor, esta diferença é o próprio conceito de erro.

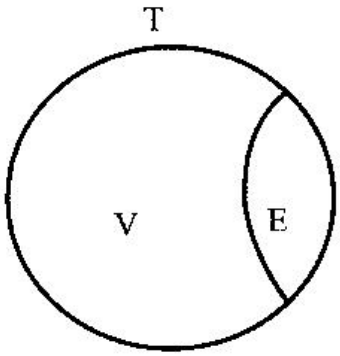


Figura 4-1. Componentes do escore T

Assim, a grande tarefa da TCT consiste em elaborar estratégias (estatísticas) para controlar ou avaliar a magnitude do E. Os erros são devidos a toda uma gama de fatores estranhos, detalhados por Campbell e Stanley (1963), tais como defeitos do próprio teste, estereótipos e vieses do sujeito, fatores históricos e ambientais aleatórios (cf. cap. 2).

Não há, contudo, maneira de se saber o escore verdadeiro do sujeito no teste, isto é, a pontuação que ele obteria se não houvesse o erro na medida. Assim, a única maneira de se obter o escore verdadeiro no teste (que é a medida não enviesada daquilo que o teste quer medir) é

aplicar este teste ao sujeito, o que irá produzir um escore empírico, o qual está contaminado pelos erros de medida. Este procedimento, contudo, produz uma equação com duas incógnitas (V e E), isto é, T = V + E, onde tanto V quanto E são desconhecidos. Mesmo que aplicássemos o teste a muitos sujeitos, cada novo sujeito entra com uma equação de duas incógnitas, e a solução se torna inviável, pois teríamos,

$$T_1 = V_1 + E_1$$

$$T_2 = V_2 + E_2$$

$$T_3 = V_3 + E_3$$

.....

$$T_i = V_i + E_i$$

onde todos os V e E são incógnitos.

Entretanto, a aplicação a muitos sujeitos irá produzir três distribuições de frequência, a saber, dos T, dos V e dos E. Então, analisando estas três distribuições talvez se possa conseguir algum progresso ou se possam fazer algumas suposições razoáveis sobre os parâmetros ou características das mesmas.

Na verdade, se consideramos o erro como um evento casualoide (cf. cap. 2), então sabemos que a média ou a expectância matemática do erro é 0, isto é,

$$\overline{E} = 0 \text{ ou } M_E = 0 \text{ ou } E(E) = 0 \qquad (4.2)$$

A média do erro é igual a 0.

Como consequência, nós temos que, pela equação 4.2, a média do escore verdadeiro será

$$\overline{V} = \overline{T} - \overline{E} = \overline{T} \text{ ou } V = E(T) - E(E) = E(T) \qquad (4.3)$$

O escore verdadeiro é a expectativa do escore empírico.

De fato,

$V = T - E$, e

$\frac{\sum V}{N} = \frac{\sum T}{N} - \frac{\sum E}{N}$, ou seja, $E(V) = E(T) - E(E) = E(T)$ ou ainda,

$V = E(T)$, dado que $E(V)$ é o próprio V.

Assim, o escore verdadeiro (V) é a esperança matemática do escore empírico (T). O conceito de esperança matemática significa o seguinte: se o sujeito respondesse infinitas vezes ao mesmo teste, ele teria infinitos diferentes escores empíricos. Estes seriam diferentes por causa do erro sempre presente nas infinitas medidas e, como o erro não é sempre da mesma magnitude (pode ser maior ou menor cada vez que o sujeito se submete ao teste, isto é, o escore empírico às vezes superestima o V e outras vezes o subestima), os escores variam a cada aplicação do teste. Contudo, a média desses infinitos escores empíricos seria o escore verdadeiro, porque ela eliminaria os erros[4], dado que uns erros aumentam e outros diminuem o escore empírico; de sorte que, no final das contas, os erros se compensariam e se eliminariam mutuamente. Isto ocorreria na suposição de que as respostas dadas ao teste pelo mesmo sujeito na ocasião 1 não afetem as dadas na ocasião 2 ou ocasião 3, e assim por diante, isto é, as respostas ao teste nas diferentes ocasiões devem ser consideradas independentes. Consequentemente, eliminado o erro, o que fica do escore empírico é o escore verdadeiro.

Além disso, podemos facilmente pressupor que não haja correlação entre o escore verdadeiro e o erro, pois não há nenhuma razão para se supor que escores verdadeiros maiores terão erros positivos e escores verdadeiros menores terão erros negativos. Se isso ocorresse, então haveria a presença de erros sistemáticos, que poderiam ser detectados e eliminados. Mas como estamos considerando erros aleatórios, então vale que

$$r_{VE} = 0 \qquad (4.4)$$

A correlação entre o escore verdadeiro e o erro é zero.

---

4. Numa situação dessa natureza, supõe-se que os erros se distribuem normalmente, com média = 0 e variância = 1.

Similarmente, não há razão de se supor que os erros cometidos num teste estejam relacionados com os cometidos em outro teste paralelo. Assim,

$$r_{E_i E_j} = 0 \qquad (4.5)$$

Não há correlação entre os erros cometidos num teste qualquer (teste i) e num teste paralelo (teste j).

Como se supõe que os erros são aleatórios, não há razão para suspeitar que eles dependam uns dos outros, o que equivaleria dizer que não seriam aleatórios; pelo contrário, em tal caso eles covariariam de um modo sistemático e bastaria descobrir a causa desta covariação para poder eliminá-los.

Falta ainda, para completar o modelo, uma conceituação do que seja *teste paralelo*. Fala-se de dois testes, T_1 e T_2, como sendo paralelos, quando eles estão medindo a mesma coisa, porém com itens (tarefas) diferentes. Matematicamente, esses testes são equivalentes se satisfizerem as duas condições seguintes:

1) os escores verdadeiros em ambos os testes são iguais ($V_1 = V_2$) ou, pelo menos, se o segundo tem a mais ou a menos uma constante *k* que se pode determinar, sendo $V_1 = V_2 + k$. Neste último caso, diz-se que os testes são "essencialmente tau-equivalentes";

2) a distribuição dos erros (variância) em ambos os testes é igual, isto é, $Var(E_1) = Var(E_2)$. Note que não dizemos que nos dois testes se cometem os mesmos erros, porque, se assim fosse, haveria uma correlação de 1,00 entre os erros cometidos nos dois testes, contradizendo o assumido na equação 4.5, onde se dizia que os erros de dois testes paralelos têm uma correlação de 0,00. O que se afirma é que os erros de um teste e do outro teste podem ter a média de 0 (equação 4.2) e uma distribuição de variabilidade também igual, ainda que os erros nos dois testes sejam individualmente diferentes.

Contudo, quando os escores verdadeiros são iguais em ambos os testes, mas as variâncias dos erros são diferentes, Lord e Novick (1968) falam que os testes são "tau-equivalentes", isto é, os testes estão medindo a mesma, mas produzindo diferente variabilidade.

Então, temos testes paralelos quando

$$V_1 = V_2 \tag{4.6}$$

Os escores verdadeiros de testes paralelos são iguais.

e

$$Var(T_1) = Var(T_2) \quad \text{ou} \quad s_1^2 = s_2^2 = s_T^2 \tag{4.7}$$

A variância de um teste é igual à variância de um teste paralelo.

Assim, podemos sumariar o modelo da Psicometria Clássica na tabela 4-1.

Tabela 4-1. Formulação do modelo da Psicometria Clássica

Modelo $T = V + E$	o escore empírico é função do escore verdadeiro mais o erro
Postulados 1) $V = E(T)$ 2) $r_{VE} = 0$ 3) $r_{E_iE_j} = 0$	o escore esperado é o escore verdadeiro não há correlação entre escore verdadeiro e o erro os erros em testes paralelos não estão correlacionados
Definição de teste paralelo	Dois testes, **i** e **j**, são paralelos se: (1) a distribuição dos seus erros tem a mesma variância, isto é, $Var(E_i) = Var(E_j)$ e (2) os escores verdadeiros de um sujeito são iguais em ambos os testes, isto é, $V_i = V_j$

Adaptado de Muñiz, 1992.

3 – Derivações do modelo

O modelo proposto e seus postulados permitem uma série de derivações (cf. a figura 4-1 para melhor visualizar as relações que seguem), que são aqui apresentadas por serem necessárias para as discussões nos capítulos seguintes.

Para caracterizar estatisticamente uma distribuição de dados qualquer são necessários dois parâmetros, a saber, a média e a variância da mesma. Além disso, para caracterizar a relação entre duas ou mais distribuições se faz necessário estabelecer a correlação e a covariância entre as mesmas. Vejamos, então, como se calculam estes parâmetros dos testes, baseados no modelo proposto pela Psicometria.

O ponto de partida são as três seguintes equações do modelo:

$T = V + E$ *O escore empírico é igual à soma do escore verdadeiro e do erro.*

$V = T - E$ *O escore verdadeiro é igual à diferença entre o escore empírico e o erro.*

$E = T - V$ *O erro é a diferença entre o escore empírico e o escore verdadeiro.*

3.1 – A média dos escores

Do E:	Por definição $M_E = 0$	(cf. equação 4.2)	
Do T:	$M_T = M_V - M_E$,	mas como $M_E = 0$	
	$M_T = M_V$	(as médias de T e V são iguais)	
Do V:	$M_V = M_T$		(4.8)

A média do escore verdadeiro é idêntica à média do escore empírico.

3.2 – A variância dos escores

Sendo a variância a média dos desvios quadráticos, onde, por exemplo, os desvios de T são $T - M_T$, os desvios quadrados $(T - M_T)^2$ e os desvios quadráticos médios $\frac{\sum(T - M_T)^2}{N}$, as variâncias dos escores serão as seguintes (onde N é o número de sujeitos):

a) Variância do T:

Sabendo que $T = V + E$, segue que

Variância do T: $\frac{\sum(T - M_T)^2}{N} = \frac{\sum[(V - M_V) + (E - M_E)]^2}{N}$

expressando os desvios médios com letras minúsculas, a saber, $t = T - M_T$, $v = V - M_V$, $e = E - M_E$, temos

$$\frac{\sum t^2}{N} = \frac{\sum (v+c)^2}{N}, \qquad \text{que efetuando dá}$$

$$\frac{\sum t^2}{N} = \frac{\sum \left(v^2 + ve + ve + e^2\right)}{N} \qquad \text{e tirando os parênteses,}$$

$$\frac{\sum t^2}{N} = \frac{\sum v^2 + \sum e^2 + 2\sum ve}{N}.$$

Como $\frac{\sum ve}{N}$ é a covariância entre o escore verdadeiro e o erro, a qual, segundo a equação 4.4 é igual a 0, a equação se reduz a $\frac{\sum t^2}{N} = \frac{\sum v^2 + \sum e^2}{N}$ que são os desvios quadráticos médios respectivamente de *V* e de *E*, isto é, as variâncias. Assim temos que

$$s_t^2 = s_v^2 + s_e^2 \tag{4.9}$$

A variância do escore empírico é a soma das variâncias do escore verdadeiro e do erro.

b) Variância do *V* e do *E:*

Da equação 4.9 segue trivialmente que

$$s_v^2 = s_t^2 - s_e^2 \tag{4.10}$$

e

$$s_e^2 = s_t^2 - s_v^2 \tag{4.11}$$

3.3 – A correlação entre os escores

a) Correlação entre o escore verdadeiro e o escore empírico:

A equação básica da correlação entre variáveis relaciona a covariância entre duas variáveis com as suas variâncias, isto é, $r_{XY} = \frac{\sum xy}{N s_X s_Y}$.

Expressa em termos psicométricos dos escores empírico e verdadeiro, esta equação fica

$$r_{TV} = \frac{\sum tv}{N s_T s_V} \qquad (4.12)$$

Substituindo a equação 4.1 na equação 4.12, onde t = v + e, temos

$$r_{TV} = \frac{\sum (v+e)v}{N s_T s_V} \qquad (4.13)$$

Tirando os parênteses, resulta que

$$r_{TV} = \frac{\sum v^2 + \sum ve}{N s_T s_V} \qquad (4.14)$$

Nesta fórmula o termo $\frac{\sum v^2}{N}$ é a variância do escore verdadeiro (s_V^2) e o termo $\frac{\sum ve}{N}$ é a covariância entre o escore verdadeiro e o erro, que segundo a equação 4.4, é zero. Assim, a equação 4.14 se reduz a

$$r_{TV} = \frac{s_v^2}{s_t s_v} \quad . \qquad (4.15)$$

ou seja,

$$r_{TV} = \frac{s_v}{s_t} \qquad (4.16)$$

A correlação entre o escore empírico e o escore verdadeiro é igual ao quociente entre o desvio padrão do escore verdadeiro e o desvio padrão do escore empírico.

b) A correlação entre o escore bruto e o erro

Partimos novamente da equação básica da correlação entre variáveis, isto é, $r_{XY} = \frac{\sum xy}{Ns_X s_Y}$, que, expressa em termos psicométricos dos escores empírico e do erro, resulta em

$$r_{TE} = \frac{\sum te}{Ns_T s_E} \tag{4.17}$$

Substituindo a equação 4.1 na equação 4.17, temos

$$r_{TE} = \frac{\sum (v+e)e}{Ns_T s_E} \tag{4.18}$$

Tirando os parênteses, resulta que

$$r_{TE} = \frac{\sum ve + \sum e^2}{Ns_T s_e} \tag{4.19}$$

Nesta fórmula o termo $\frac{\sum e^2}{N}$ é a variância do erro (s_E^2) e o termo $\frac{\sum ve}{N}$ é a covariância entre V e E, a qual é zero, resultando que a equação 4.19 fica

$$r_{TE} = \frac{s_e^2}{s_t s_e} \tag{4.20}$$

ou seja,

$$r_{TE} = \frac{s_e}{s_t} \tag{4.21}$$

A correlação entre o escore empírico e o erro é igual ao quociente entre o desvio padrão do erro e o desvio padrão do escore empírico.

Além dessas equivalências, são verdadeiras, pelo modelo psicométrico, também as seguintes:

$Cov(V,E) = 0$	Não há covariância entre o escore verdadeiro e os erros.	(4.22)
$Cov(T,V) = Var(V)$	A covariância entre o escore empírico e o escore verdadeiro é a variância do escore verdadeiro (todo o escore V está incluído no escore T); é a interseção $T \cap V$, onde $V \subset T$, isto é, V está incluído em T.	(4.23)
$Cov(T_i,T_j) = Cov(V_i,V_j)$	A covariância entre escores empíricos de dois testes é igual à covariância entre os seus escores verdadeiros	(4.24)
$M_1 = M_2 = ... = M_i$	Para k testes paralelos, as médias,	(4.25)
$s_1^2 = s_2^2 = ... = s_i^2$	as variâncias	(4.26)
$r_{12} = r_{13} = ... = r_{ij}$	e as correlações são todas iguais entre si (por postulados de teste paralelo).	(4.27)

Embora a teoria clássica assuma que as estatísticas baseadas em amostras suficientemente grandes constituem estimativas adequadas dos parâmetros da população (as relações feitas acima nas derivações se referem a estes parâmetros), as três últimas derivações (média, variância e correlação) podem ser testadas empiricamente. As outras derivações, na verdade, constituem apenas tautologias dentro do modelo (cf. Lord & Novick, 1968).

II – O Modelo da Psicometria Moderna: TRI

1 – A Teoria de Resposta ao Item e a teoria psicométrica clássica

A Teoria de Resposta ao Item (TRI) já tem uma longa história. Ela se baseia no modelo do traço latente, que possui uma história mais longa ainda, dentro do qual você encontra autores dos anos de 1930, tais como Thurstone (Lumsden, 1980), Likert (Andersen, 1977; Andrich, 1978), Lawley (1943), Guttman (1941, 1944), Lazarsfeld (1950). Contudo, a TRI

começou a ser formalizada mais tecnicamente com os trabalhos de Lord (1952, 1953) nos Estados Unidos e Rasch (1960) na Dinamarca, que a utilizaram para testes de desempenho e de aptidão. Entretanto, apenas ultimamente, a partir de meados dos anos de 1980, a TRI vem se tornando a técnica predominante no campo dos testes. A razão da demora desta teoria em ser amplamente utilizada em Psicometria se deve ao fato da enorme complexidade de manipulação de seus modelos matemáticos, inviáveis sem complexos programas de computador e estes só começaram efetivamente a entrar no mercado nos anos de 1980, com os procedimentos de estimação dos parâmetros do modelo desenvolvidos por Wood, Wingersky e Lord (1978) e por Wingersky, Barton e Lord (1982).

Atualmente, a TRI parece que veio para ficar e substituir grande parte da teoria clássica da Psicometria. Isto é um fato em países do Primeiro Mundo (Estados Unidos, Canadá, Europa, Japão, Israel, Austrália); no restante do mundo ela é ainda raramente utilizada e no Brasil (América Latina em geral) ela chegou a ser conhecida apenas nos anos de 1990. Este capítulo e os subsequentes visam tornar a teoria e a técnica da TRI conhecida e utilizada por um número maior de pesquisadores no país.

As publicações em TRI vêm crescendo e tomando conta das revistas especializadas, como a *Psychometrika*. Há centros importantes de pesquisa nesta área nos Estados Unidos (*University of Massachusetts* em Amherst), Holanda e Espanha (Universidade de Oviedo). Existe, inclusive, uma sociedade internacional, a *International Test Commission* (ITC), que filia seguidores da TRI. De fato, já no Congresso Internacional da ITC, em Oxford (Inglaterra), de julho de 1993, havia mais de 120 participantes de cerca de 46 países. Da América Latina só estavam presentes o Brasil e a Argentina com dois representantes cada.

O enorme impacto que a TRI vem tendo em Psicometria se deve ao fato dela superar certas limitações teóricas graves que a psicometria tradicional contém. Hambleton, Swaminathan e Rogers (1991) salientam especialmente quatro dessas limitações:

1) Os parâmetros clássicos dos itens (dificuldade e discriminação) dependem diretamente da amostra de sujeitos utilizada para estabelecê-los (*group-dependent* ou *sample-dependent*). Daí, se a amostra não for rigorosamente representativa da população, aqueles parâmetros dos itens não podem ser considerados válidos para esta população. Como conseguir amostras representa-

tivas é um problema prático grave para os construtores de testes, a dependência dos parâmetros dos itens na amostra obtida se torna um empecilho de grandes proporções para a elaboração de instrumentos psicométricos não enviesados.

2) A avaliação das aptidões dos testandos também depende do teste utilizado (*test-dependent*). Assim, testes diferentes que medem a mesma aptidão irão produzir escores diferentes da mesma aptidão para sujeitos idênticos. Testes com índices de dificuldade diferentes evidentemente produzirão escores diferentes. No caso das formas paralelas de testes, é preciso observar que, em primeiro lugar, conseguir formas estritamente paralelas é uma tarefa quase impossível e, em segundo lugar, mesmo conseguindo formas paralelas, é difícil pressupor que elas produzem o mesmo montante de erro, o que vem a afetar a estimação do escore verdadeiro dos sujeitos.

3) A definição do conceito de fidedignidade ou precisão na teoria clássica dos testes constitui também uma fonte de dificuldades. Ela é concebida como a correlação entre escores obtidos de formas paralelas de um teste ou, mais genericamente, como o oposto do erro de medida. Ambos os conceitos apresentam dificuldades. Em primeiro lugar, é praticamente impossível satisfazer as condições de definição de formas paralelas e, no caso do erro de medida, é postulado que este seja idêntico em todos os examinandos, postulado improvável (Lord, 1984), uma vez que fica difícil presumir que sujeitos de baixa aptidão, por exemplo, cometam erros iguais aos de habilidades superiores.

4) Outro problema da teoria clássica dos testes consiste em que ela é orientada para o teste total e não para o item individual. Toda a informação do item deriva de considerações do teste geral, não se podendo assim determinar como o examinando se comportaria diante de cada item individual. Ademais, a análise de cada item é feita em função do escore total, do qual cada item faz parte. Então, fica um tanto incongruente avaliar a qualidade do item quando ele próprio contribui para a mesma e, ainda, a admissão de um escore total já supõe que os itens sejam adequados; em sendo o caso, para que fazer a análise individual de

cada item em função de todos os outros, que, aliás, ainda não foram analisados em sua adequação?

Estas e outras dificuldades dos modelos e técnicas clássicos de medida incitaram os psicometristas à procura de teorias alternativas que pudessem permitir estabelecer (Hambleton et al., 1991):

1) características do item sem ser dependentes da amostra de sujeitos utilizados;

2) escores dos examinandos independentes do teste utilizado;

3) um modelo ao nível do item em vez do teste, de sorte que a análise do item não dependa dos demais itens do teste;

4) um modelo que não exija formas rigorosamente paralelas para avaliar a fidedignidade;

5) um modelo que ofereça uma medida de precisão para cada nível de aptidão, isto é, que a avaliação da aptidão tenha igual exatidão em todos os seus níveis e não somente nos níveis medianos como faz a psicometria clássica.

Essas características são precisamente oferecidas pela teoria da resposta ao item (Hambleton, 1983; Hambleton & Swaminathan, 1985; Lord, 1980; Wright & Stone, 1979; Hambleton, Swaminathan & Rogers, 1991; Muñiz, 1990).

2 – Características da TRI

2.1 – A teoria da TRI

Diferentemente da teoria clássica de Psicometria, a TRI trabalha com traços latentes e adota dois axiomas fundamentais:

1) o desempenho do sujeito numa tarefa (item do teste) se explica em função de um conjunto de fatores ou traços latentes (aptidões, habilidades, etc.). O desempenho é o efeito e os traços latentes são a causa;

2) a relação entre o desempenho na tarefa e o conjunto dos traços latentes pode ser descrita por uma equação monotônica crescente, chamada de CCI (Função Característica do Item ou Curva Característica do Item) e exemplificada na figura 4-2, onde se observa

que sujeitos com aptidão maior terão maior probabilidade de responder corretamente ao item e vice-versa (θ_i é a aptidão e $P_i(\theta)$ a probabilidade de resposta correta dada ao item).

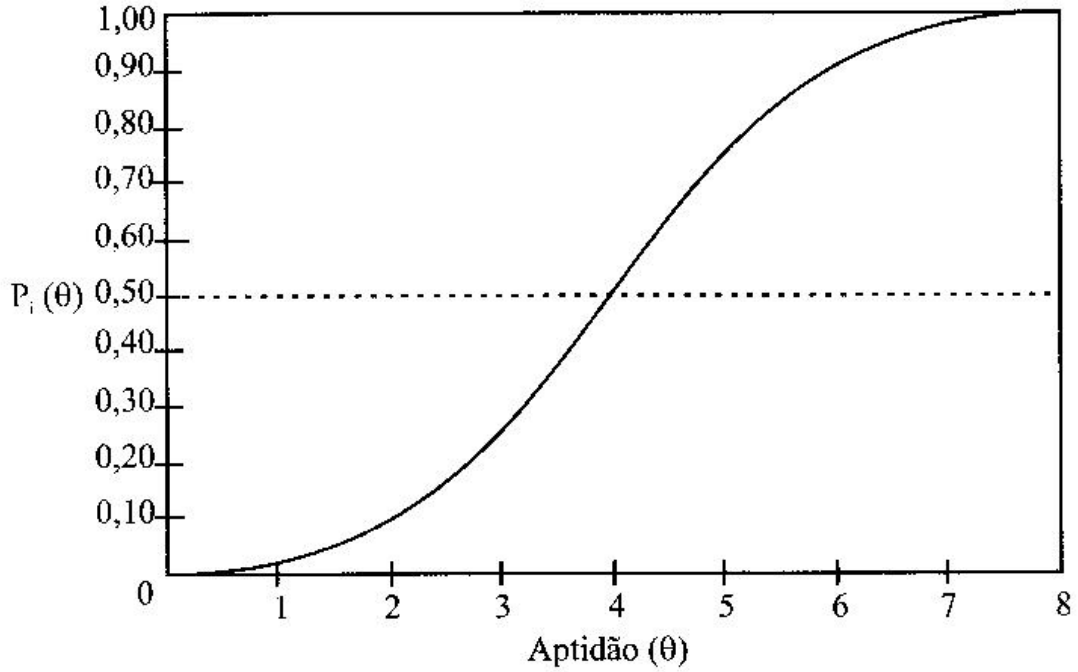


Figura 4-2. A curva CCI

Concretamente, a TRI está dizendo o seguinte: Você apresenta ao sujeito um estímulo ou uma série de estímulos (tais como itens de um teste) e ele responde aos mesmos. A partir das respostas dadas pelo sujeito, isto é, analisando as suas respostas aos itens especificados, pode-se inferir sobre o traço latente do sujeito, hipotetizando relações entre as respostas observadas deste sujeito com o nível do seu traço latente. Estas relações podem ser expressas através de uma equação matemática que descreve a forma de função que estas relações assumem.

De fato, pode-se imaginar um número ilimitado de modelos matemáticos que podem expressar esta relação, dependendo do tipo de função matemática utilizada e/ou do número de parâmetros que se quer descobrir para o item. Logo mais voltaremos a este ponto, mas antes anote uma preciosa vantagem, sobre a teoria clássica, que a TRI tem quanto aos modelos que usa, a saber, os modelos utilizados pela TRI permitem desconfirmação. Na verdade, a demonstração da adequação do modelo aos dados (*model-data goodness-of-fit*) é um passo necessário nos procedimentos desta teoria.

2.2 – *Pressupostos da TRI*

Entre as suposições que a TRI faz, duas são de especial relevância e precisam ser descritas: a unidimensionalidade e a independência local.

Unidimensionalidade

A TRI postula que há apenas uma aptidão responsável pela (ou traço subjacente à) realização de um conjunto de tarefas (itens). Entre os psicólogos, é pacífico que qualquer desempenho humano é sempre multideterminado ou multimotivado, dado que mais de um traço latente entra na execução de qualquer tarefa. Contudo, para satisfazer o postulado da unidimensionalidade é suficiente admitir que haja uma aptidão *dominante* (um fator dominante) responsável pelo conjunto de itens. Este fator é o que se supõe estar sendo medido pelo teste. Essa questão da unidimensionalidade de um conjunto de itens (leia teste) está ainda produzindo muita dor de cabeça para os pesquisadores da TRI. De fato, há tentativas de trabalhar os modelos logísticos (que são os modelos hoje predominantes na TRI[5]), levando em conta o fato da multideterminação nas respostas dadas a um teste; são os modelos ditos da Teoria Multidimensional de Resposta ao Item (MIRT ou MTRI – Reckcase, 1985; Reckcase & McKinley, 1991; Reckcase, 1994). Contudo a TRI unidimensional é de longe a mais comumente utilizada no estudo dos testes e, assim, o postulado da unidimensionalidade ainda continua importante, especialmente porque soluções para modelos multidimensionais levantam ainda muita celeuma. É, entretanto, importante que você esteja consciente do problema da multideterminação do comportamento humano e que soluções multidimensionais parecem ser mais condizentes com a realidade. Mesmo assim, a solução unidimensional que a TRI traz é considerada bastante robusta, isto é, os desvios que traços latentes secundários, além do traço dominante, produzem na interpretação dos escores de um teste são suficientemente pequenos para poderem ser negligenciados (Junker & Stout, 1994; McDonald, 1994; Gessaroli, 1994).

---

5. Modelos logísticos são modelos que trabalham com logaritmos.

Independência local

Este postulado afirma que, mantidas constantes as aptidões que afetam o teste, as respostas dos sujeitos a quaisquer dos itens são estatisticamente independentes. Isto quer dizer que os itens são respondidos em função do traço latente predominante e não em função de memória ou outros traços latentes.

Para melhor entender esta história da independência local, considere o seguinte: Seja

- θ o conjunto de aptidões que afetam um conjunto de itens
- U_i a resposta de um sujeito ao item *i* (i = 1, 2, ..., n) e
- $P(U_i \mid \theta)$ a probabilidade de resposta do sujeito *j* com aptidão θ (lê-se: probabilidade de resposta ao item, dado um certo θ). Assim, $P(U_i=1 \mid \theta)$ significa a probabilidade de uma resposta correta (isto é, vale 1) e $P(U_i=0 \mid \theta)$ a probabilidade de uma resposta errada (isto é, valendo 0).

Com essas informações, a independência local pode ser matematicamente afirmada como

$$\text{Prob}(U_1, U_2, ..., U_n \mid \theta) = P(U_1 \mid \theta)\, P(U_2 \mid \theta) ..... P(U_n \mid \theta) \qquad (4.28)$$

$$= \prod_{i=1}^{n} P(U_i \mid \theta).$$

Assim, a independência local significa que, para examinandos com uma aptidão dada, a probabilidade de resposta a um conjunto de itens é igual aos produtos das probabilidades das respostas (produtório) do examinando a cada item individual. Assim, se um sujeito acertou os itens 1 e 2 e errou o 3, a configuração ou o padrão de suas respostas é $U_1=1$, $U_2=1$, $U_3=0$, ou seja, 1 1 0, e a independência local implica que

$$P(U_1=1, U_2=1, U_3=0 \mid \theta) = P(U_1=1 \mid \theta)\, P(U_2=1 \mid \theta)\, P(U_3=0 \mid \theta)$$

$$= P_1 P_2 Q_3$$

sendo $P(U_i = 1 \mid \theta)$ e $Q_i = 1 - P_i$.

Embora pareça improvável que os comportamentos de um mesmo sujeito não estejam correlacionados, a independência local afirma que, se houver correlação, esta se deve à influência de fatores outros (secundários) que não o fator dominante. Se estes outros fatores forem controlados (mantidos constantes), o fator dominante será a única fonte de variação e as respostas se tornam independentes. Assim, a independência local das respostas dos sujeitos implica também a unidimensionalidade do teste como acima definido (Lord, 1980; Lord & Novick, 1968), isto é, que os itens que estão sendo analisados estejam medindo o mesmo traço latente dominante, em função do qual as respostas a cada item são dadas.

3 – Modelos da TRI

Embora seja ilimitado o número de modelos matemáticos que podem expressar a relação de probabilidade de sucesso em um item e a aptidão medida pelo teste (isto é, a CCI), na prática há três que predominam. Estes se distinguem pelo número de parâmetros que utilizam para descrever o item, a saber: os modelos logísticos de 1, 2 e 3 parâmetros, isto é, modelos que avaliam somente a dificuldade do item, ou a dificuldade e a discriminação, ou a dificuldade, a discriminação e a resposta correta dada ao acaso.

3.1 – Modelo logístico de 1 parâmetro

Este modelo, inicialmente criado por Rasch (1960) e expresso como modelo de ogiva, foi descrito para um modelo logístico por Wright (1977a, 1977b), o qual permite tratamento matemático mais fácil. Sua fórmula é

$$P_i(\theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}} \qquad (i = 1, 2, ..., n) \qquad (4.29)$$

onde

$P_i(\theta)$ é a probabilidade de um examinando com aptidão θ responder o item *i* e é representado como uma curva tipo S

b_i é o parâmetro de dificuldade do item *i*

n é o número de itens no teste

e é um número transcendental com valor de 2,7182818... (= 2,72)

D é uma constante que vale 1,7.

$P_i(\theta)$ produz uma curva, chamada curva característica do item (CCI – *Item Characteristic Curve*), conforme a figura 4-2 (acima). O parâmetro de dificuldade b_i do item corresponde ao ponto na escala de aptidão θ onde a probabilidade de resposta é 0,50. Quanto maior for o b_i, maior deve ser o nível de aptidão exigido para que o examinando tenha a chance de 50% de acertar o item. Transformando a escala da aptidão em escores padrões, com média = 0 e desvio padrão = 1, os valores de b_i tipicamente se situam entre -3 (itens fáceis) e +3 (itens difíceis), conforme figura 4-3, onde o item 1 exige aptidão de cerca de 0 e o item 2 aptidão de 0,95, sendo este item mais difícil que o item 1.

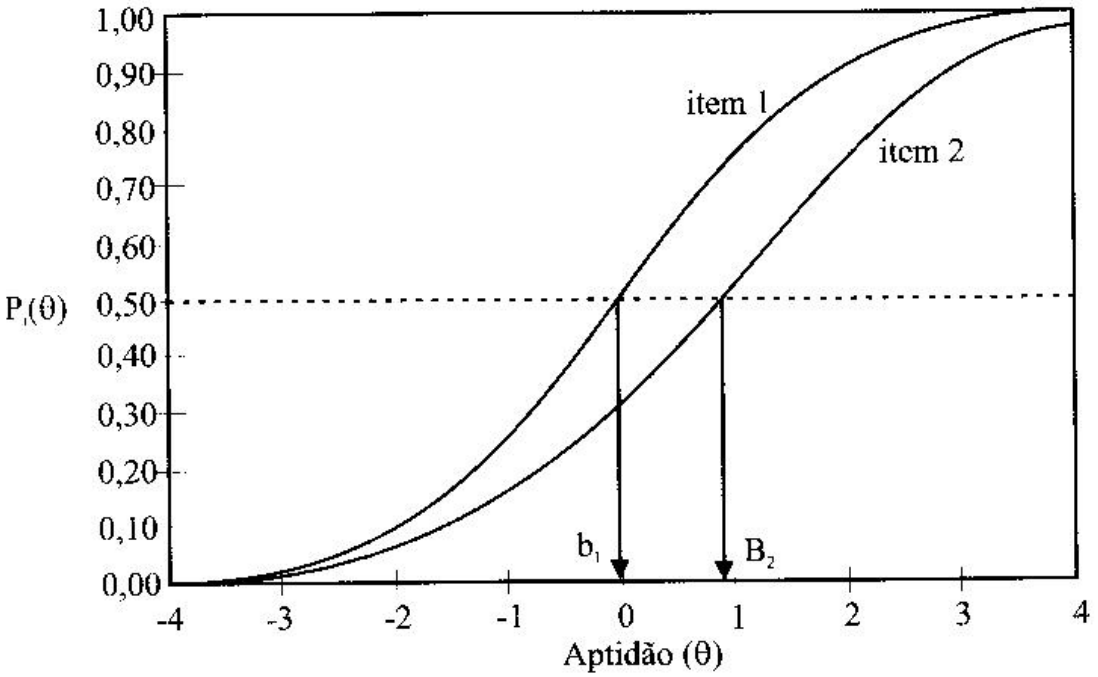


Figura 4-3. Parâmetro de dificuldade (b) de dois itens

A constante *D* foi incluída na fórmula para tornar a curva logística, com a qual trabalha a TRI moderna, igual à curva normal acumulada (ogiva) utilizada nos estudos pioneiros da TRI.

3.2 – Modelo logístico de 2 parâmetros

Birnbaum (1968) desenvolveu a equação que serve para avaliar dois parâmetros do item: dificuldade e discriminação. A fórmula é

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad (4.30)$$

onde a_i é o parâmetro de discriminação do item, que pode variar de 0 a ∞, mas tipicamente varia entre 0 e 2. Valores negativos indicariam que a probabilidade de acertar um item estaria inversamente relacionada com a aptidão, o que soa estranho, porque indicaria que o item é corretamente acertado por sujeitos de menor habilidade e errado pelos de maior habilidade. A figura 4-4 mostra os dois parâmetros do item (b_i e a_i). O a_i é representado pela inclinação da curva (ângulo) no ponto de inflexão, onde a probabilidade de resposta correta é 0,50.

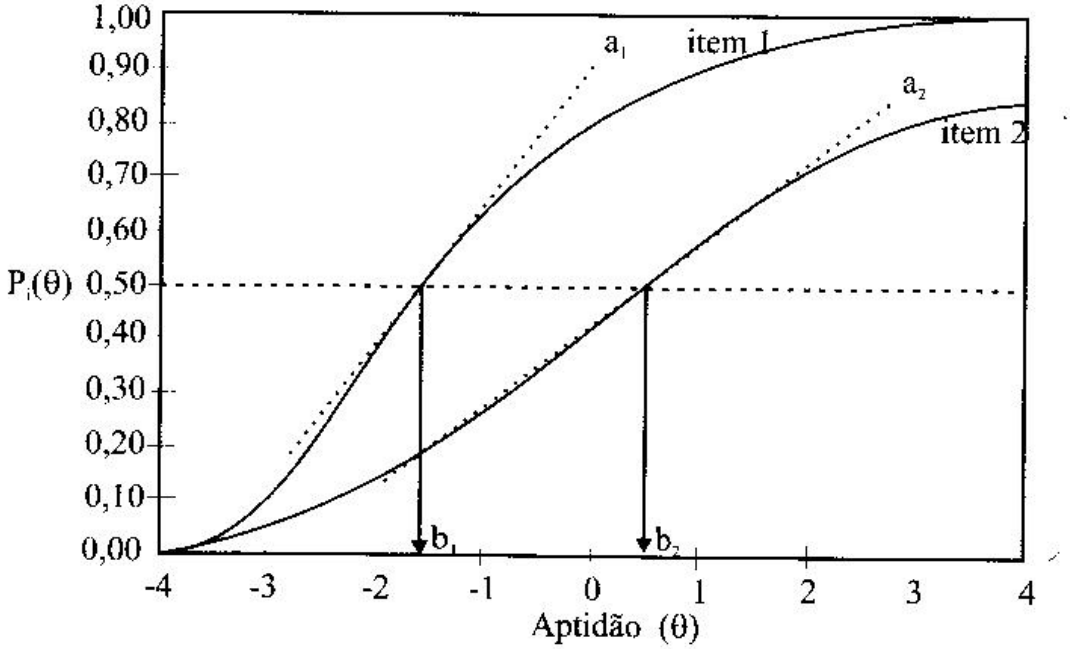


Figura 4-4. Parâmetros de dificuldade (b) e discriminação (a) de dois itens

Na ilustração, o item 2 (parâmetro b_2) é mais difícil que o item 1 (parâmetro b_1), mas menos discriminativo (parâmetro a_2) pois a inclinação da curva dele é menor que a do item 1 (parâmetro a_1).

3.3 – Modelo logístico de 3 parâmetros

Desenvolvida por Lord (1980), a fórmula deste modelo é:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta\ b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad (4.31)$$

onde, c_i é o parâmetro do item que avalia a resposta correta dada ao item por acaso e é expresso pela assíntota inferior da curva. Se esta assíntota cortar a ordenada acima do ponto 0, há presença de acertos ao acaso (figura 4-5). No caso do item 2 na figura 4-5, há 20% de probabilidade que o item seja acertado por acaso, sendo esta probabilidade de 0 para os outros dois itens. A lógica que fundamenta essa interpretação da assíntota é a seguinte: supostamente o sujeito não tem habilidade praticamente nenhuma, pois ele tem um θ menor que –3, e apesar disso acerta o item; consequentemente, ele só pode ter chutado e teve sorte, porque acertou.

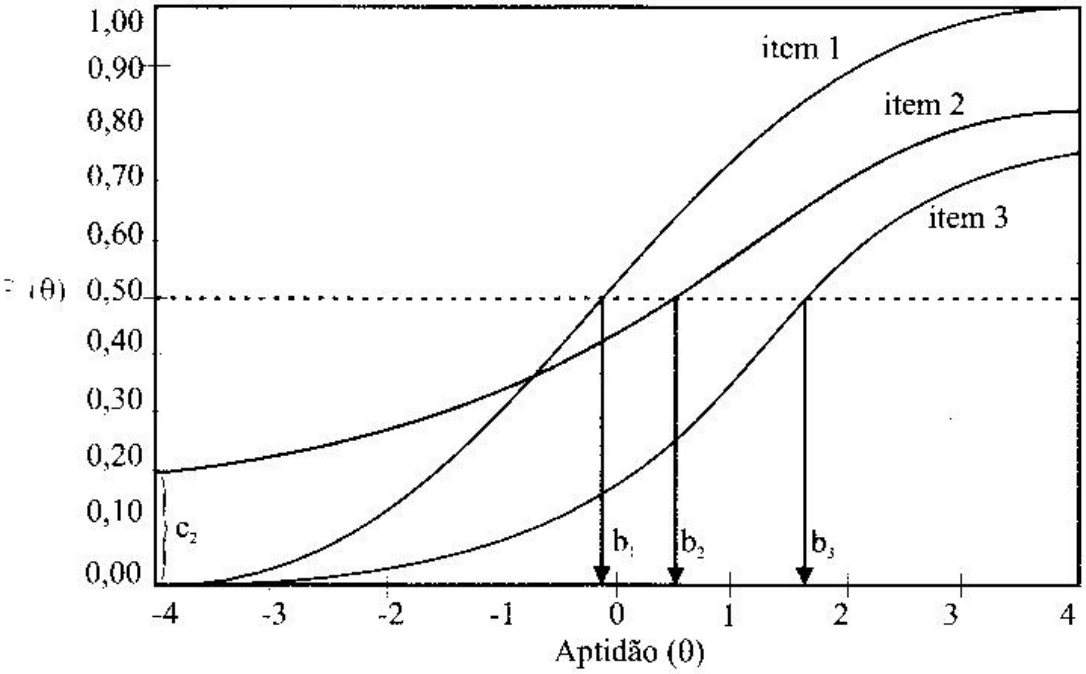


Figura 4-5. CCI do modelo de três parâmetros para três itens

Veja os exemplos da tabela 4-2 para sentir como essa fórmula funciona para determinar a probabilidade de acertar um item, variando os parâmetros dos itens e do teta. Observe como foi calculada a probabilidade de acerto do item 6 da tabela pelo modelo logístico de 3-parâmetros:

$$P_6(\theta) = c_6 + (1 - c_6)\frac{e^{Da_6(\theta\ b_6)}}{1 + e^{Da_6(\theta\ b_6)}}$$

$$= 0{,}25 + (1 - 0{,}25)\frac{2{,}7183^{1{,}7\times 1{,}8(2{,}05-2{,}1)}}{1 + 2{,}7183^{1{,}7\times 1{,}8(2{,}05-2{,}1)}} = 0{,}60$$

Tabela 4-2. Probabilidade de acertar um item variando os parâmetros dos itens e do teta

Item	Parâmetros				$P_i(\theta)$
	a	b	c	θ	
1	1,00	-3,00	0,00	1,00	1,00
2	1,50	-2,10	0,10	-2,10	0,55
3	2,00	-0,50	0,15	-1,30	0,20
4	2,50	0,00	0,20	0,50	0,91
5	2,10	1,30	0,11	1,00	0,34
6	1,80	2,10	0,25	2,05	0,60
7	1,50	3,00	0,16	2,50	0,34

Você vê nessa tabela que o item 1 é extremamente fácil (b = -3,00) e o sujeito tem aptidão acima da média (θ = 1,00), tendo como consequência uma probabilidade de 100% de acertar o item. Ao passo que o item 7, que é extremamente difícil (b = 3,00) e o sujeito que responde é extremamente inteligente (θ = 2,50), tem uma probabilidade de apenas 34% de ser acertado, porque o item é mais difícil do que o tamanho da aptidão do sujeito; e assim por diante.

4 – Determinação dos parâmetros de itens e de aptidões

A determinação destes parâmetros constitui apenas uma das etapas na elaboração de instrumentos psicológicos. As etapas da elaboração de instrumentos dentro da TRI dividem-se em três níveis de procedimentos[6]:

1) Procedimentos teóricos onde se incluem as etapas de

a) estabelecimento do sistema ou variável (traço latente) a ser medido;

---

6. Uma exposição sobre estes passos pode ser encontrada no livro deste autor "Testes psicológicos: Manual prático de elaboração", cap. 3. Brasília, DF: LabPAM/IBAPP.

b) desenvolvimento da teoria psicológica sobre este traço;

c) operacionalização do traço através da elaboração dos comportamentos que o representam (elaboração dos itens) e

d) análise teórica dos itens.

2) Procedimentos empíricos, que consistem em

a) definição da amostra de sujeitos para a coleta da informação sobre o teste que se quer utilizar no futuro na população e

b) aplicação dos itens a esta amostra.

3) Procedimentos analíticos, que consistem em

a) escolha do modelo de TRI;

b) estabelecimento da dimensionalidade do traço (unidimensionalidade dos itens);

c) avaliação dos parâmetros dos itens e da aptidão do sujeito (o traço θ) e

d) demonstração da adequação do modelo aos dados empíricos.

Na TRI, o desempenho do sujeito numa tarefa (item), isto é, a probabilidade de resposta correta [$P_i(\theta)$] depende de: (1) aptidão do sujeito (θ) e (2) dos parâmetros dos itens (a_i, b_i e c_i). Daí, a primeira tarefa da TRI é viabilizar modelos que possam permitir a descoberta dos parâmetros dos itens. As fórmulas dadas anteriormente para os modelos da TRI são equações, onde ocorrem constantes (como o *e* e o *D*) e também variáveis, como os parâmetros dos itens (a, b, c) e a aptidão dos sujeitos (o teta). A tarefa da TRI consiste em estimar os valores destes parâmetros, tais que melhor expliquem os resultados obtidos. Esta estimação é feita com base nos dados empíricos, isto é, as respostas da amostra de sujeitos aos itens. Isto consiste em se escolher como parâmetros para os itens aqueles valores que maximizam a probabilidade de ocorrência dos dados que de fato apareceram nas respostas dos sujeitos. Por exemplo, se ao lançar 100 vezes uma moeda (cara ou coroa) e aparecer 60 caras e 40 coroas, a probabilidade mais verossímil de que apareça cara é de 60/100, isto é, 0,60. Assim, estima-se que 0,60 é o valor mais provável de aparecer cara. Este método

de avaliação se chama de máxima verossimilhança (*maximum likelihood*)[7], porque os valores estimados são os mais verossímeis, plausíveis, com respeito aos dados empíricos obtidos.

A estimação dos parâmetros se faz normalmente em dois passos:

1) estimação dos parâmetros de cada item (isto é, os parâmetros *a*, *b*, *c*), chamada também de calibração ou parametrização. Há várias maneiras para proceder a esta estimação, tais como os procedimentos *Ancilles* e *Logist* (Swaminathan & Gifford, 1994);

2) estimação dos níveis do traço latente (o teta) dos sujeitos, utilizando os parâmetros dos itens agora já conhecidos, isto é, estes parâmetros já se tornaram constantes na equação.

A estimação de todos estes parâmetros se faz por aproximações sucessivas (iteração), utilizando-se pacotes estatísticos apropriados, tais como BICAL (Wright et al., 1979) para modelos logísticos de 1 parâmetro e BILOG (Mislevy & Bock, 1984) e LOGIST (Wingersky, Barton & Lord, 1982) para modelos de 1, 2 e 3 parâmetros[8]. Estes pacotes produzem tanto os parâmetros dos itens quanto os valores de θ dos sujeitos (Baker, 1987; Birnbaum, 1968; Lord, 1980; Swaminathan, 1983; Hambleton & Swaminathan, 1985).

A lógica da estimação dos parâmetros na TRI é mais ou menos a seguinte (Lord, 1980): parte-se de uma matriz composta de *n* itens (i) com as respostas a eles dadas por N sujeitos (j), isto é, uma matriz $U_{ij} = [u_{ij}]$. Nesta matriz, cada sujeito aparece com o seu padrão de resposta conforme tabela 4-3.

---

7. Uma exposição sobre este método matemático se encontra no livro deste autor "Análise fatorial para pesquisadores", Petrópolis, RJ, Editora Vozes, s.d.

8. Notas sobre pacotes estatísticos para a TRI se encontram no Apêndice C.

Tabela 4-3. Matriz U_{ij}

	Item					
Sujeito	1	2	3	4	...	n
1	1	1	0	1	...	0
2	1	0	1	1	...	1
3	0	1	1	0	...	0
4	1	1	1	0	...	0
5	1	1	0	0	...	1
...	...	...	...	...	...	...
N	1	1	0	1	...	1

A suposição é de que esta matriz surgiu de um modelo como o da fórmula 4.31. O que é preciso fazer, a partir dessa matriz, consiste em estimar os parâmetros dos itens (a, b, c) e o teta (θ). Para cada um desses parâmetros temos uma equação de verossimilhança para resolver; essas equações são as seguintes:

$$\sum_{i=1}^{n} \frac{u_{ij} - P_{ij}}{P_{ij}Q_{ij}} \frac{\partial P_{ij}}{\partial \theta_j} = 0 \qquad (j = 1, 2, \cdots, N)$$

$$\sum_{j=1}^{N} \frac{u_{ij} - P_{ij}}{P_{ij}Q_{ij}} \frac{\partial P_{ij}}{\partial a_i} = 0 \qquad (i = 1, 2, \cdots, n)$$

$$\sum_{j=1}^{N} \frac{u_{ij} - P_{ij}}{P_{ij}Q_{ij}} \frac{\partial P_{ij}}{\partial b_i} = 0$$

$$\sum_{j=1}^{N} \frac{u_{ij} - P_{ij}}{P_{ij}Q_{ij}} \frac{\partial P_{ij}}{\partial c_i} = 0$$

Não se assuste com estas fórmulas; elas estão aí apenas para mostrar que para resolvê-las você vai primeiro descobrir o θ, utilizando a primeira das equações acima e assumindo valores para os outros parâmetros (a, b, c). Descoberto o θ dos sujeitos, então você vai estimar os parâmetros dos itens, resolvendo as três últimas equações; para que os valores con-

virjam, é preciso repetir (iterar) o processo inúmeras vezes, donde a estória da iteração na estimação dos parâmetros na TRI.

Vamos ilustrar brevemente a lógica desses procedimentos, seguindo Muñiz Fernández (1990). Os valores estimados para os parâmetros pelo método da verossimilhança são os que maximizam a probabilidade de que ocorram aqueles valores (respostas) dados pelos sujeitos. Então procura-se uma função matemática que produza estes máximos (que é a derivada expressa nas fórmulas pelo ∂), como ilustrado na figura 4-6 (onde o sujeito 1 tem seu máximo em $\theta = -1,95$ e o sujeito 2 em $\theta = 0,95$).

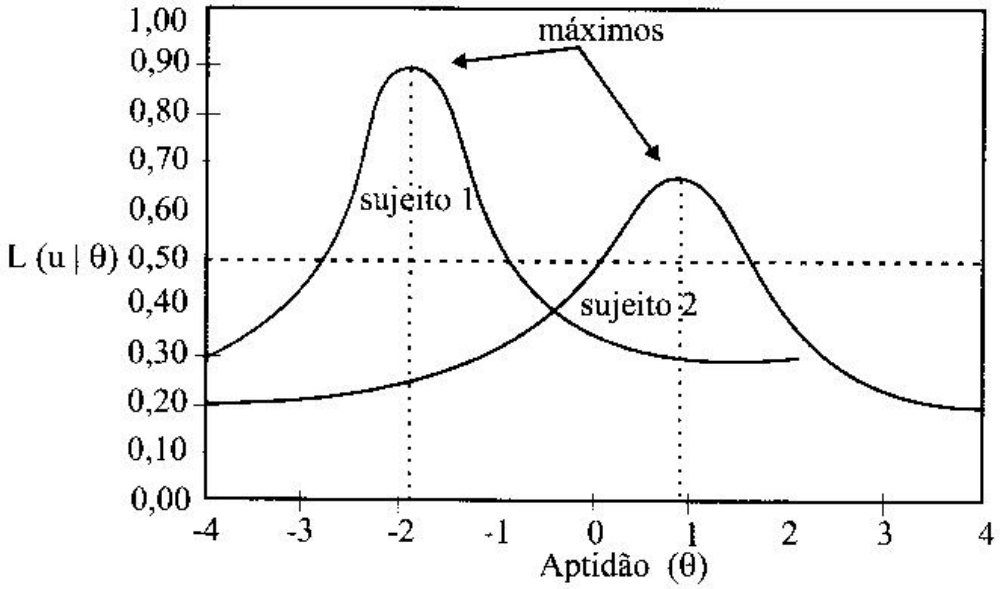


Figura 4-6. Máximos da função de verossimilhança

Seja U_i a resposta ao item, onde, em testes de aptidão, $U_i = 1$ se o item for corretamente respondido e $U_i = 0$ se o sujeito errar o item. A equação CCI dá precisamente a probabilidade de acerto e erro para um dado valor de aptidão θ, tal que $P(U_i=1 \mid \theta) = P(\theta)$ e $P(U_i=0 \mid \theta) = 1 - P(\theta) = Q(\theta)$, onde se entende $P(U_i=1 \mid \theta)$ como a probabilidade de uma resposta correta para um tal valor de θ. Exemplo: a probabilidade de acertar o item na figura 4-7 é de 0,60 e de errar é de 0,40 para um valor θ de 1, ou seja,

$$P(U_i=1 \mid \theta = 1) = 0,60$$

$$P(U_i=0 \mid \theta = 1) = 0,40.$$

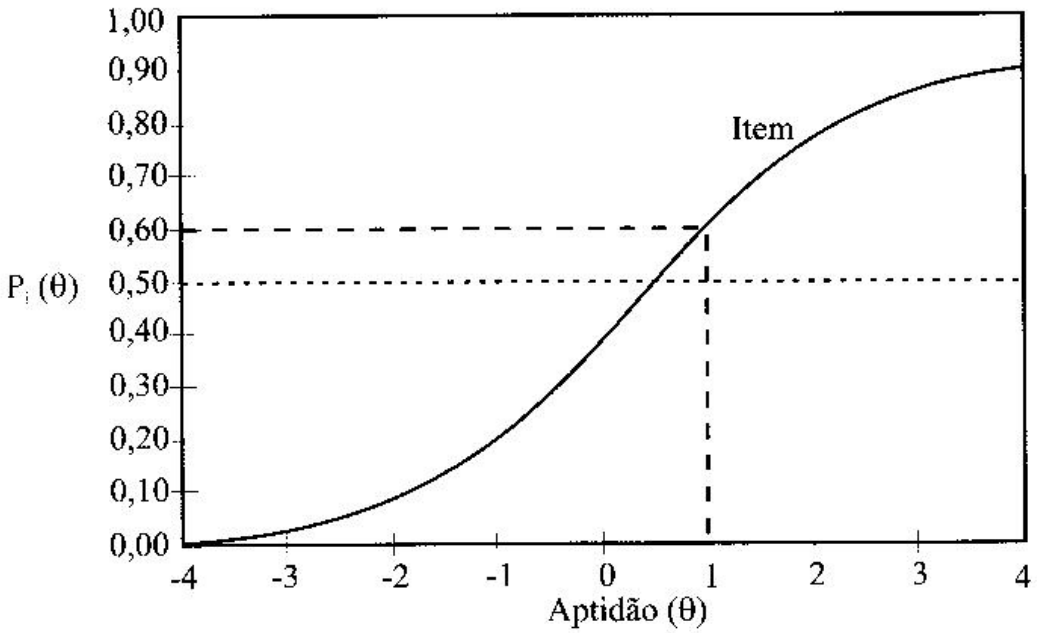


Figura 4-7. Probabilidade 0,60 de acerto do item a $\theta = 1$

Já que a resposta a um item para um dado valor de θ é considerada uma prova de Bernoulli[9], segue que $P(U_i|\theta) = P(U_i = 1|\theta)^{U_i} P(U_i = 0|\theta)^{(1-U_i)}$

$$=[P(\theta)]^{U_i} [Q(\theta)]^{(1-U_i)}.$$

No nosso caso

$$P(U_1=1 \mid \theta = 1) = (0,60)^1 (0,40)^{(1-1)} = 0,60$$
$$P(U_1=0 \mid \theta = 1) = (0,60)^0 (0,40)^{(1-0)} = 0,40$$

Como um teste tem *n* itens, a probabilidade de um padrão de resposta é dada pelo produto das probabilidades de cada item (dado o axioma da independência local). Assim, um padrão 11010 para cinco itens será:

$$P_1(\theta)P_2(\theta)Q_3(\theta)P_4(\theta)Q_5(\theta)$$

---

9. Esta prova de Bernoulli significa o seguinte: são experimentos independentes repetidos que possuem apenas duas possibilidades de resultados para cada experimento, como é o caso de um item ser acertado ou errado ou o lançar de uma moeda, e cuja probabilidade para cada resultado se mantém constante durante todo o experimento, e que está sintetizada no pressuposto da unidimensionalidade dos itens.

ou, em geral,

$$L(u_1, u_2, \ldots, u_n \mid \theta) = \prod_{i=1}^{n} [P_i(\theta)]^{U_i} [Q_i(\theta)]^{(1-U_i)} \qquad (4.32)$$

que é a função de verossimilhança (*likelihood function*).

Esta equação vem normalmente expressa em termos de logaritmos[10] por ser mais fácil de matematicamente operar, pois

$\ln xy = \ln x + \ln y$ (logaritmo natural do produto xy é igual à soma dos logaritmos de x e de y)

e

$\ln x^a = a \ln x$ (logaritmo de uma variável exponenciada x é igual ao produto do expoente vezes o logaritmo da variável x).

Assim, a equação de verossimilhança da TRI, em termos logísticos, se escreve como

$$\ln L(u \mid \theta) = \sum_{i=1}^{n} [U_i \ln P_i(\theta) + (1 - U_i) \ln Q_i(\theta)] \qquad (4.33)$$

onde,

ln = logaritmo natural

u = vetor das respostas.

5 – Ajuste do modelo (*Model – Data Goodness-of-Fit*)

A TRI constitui um poderoso instrumento para avaliar instrumentos psicológicos. Contudo, isto é verdade somente se o modelo for adequado para os dados empíricos coletados. Felizmente a TRI possui técnicas para verificar tal suposição, ainda que estas técnicas deixem a desejar, segundo os pesquisadores da área. Assim, para poder usufruir das vantagens da TRI, é preciso demonstrar que o modelo escolhido se adequa aos dados empíricos, isto é, se os valores $P(\theta)$ estimados pelo modelo não diferem dos valores obtidos empiricamente (a saber, a proporção de sujeitos que de fato acertaram o item).

---

10. Donde a razão dos modelos da TRI se chamarem modelos logísticos.

Para tal demonstração há uma série de procedimentos estatísticos que constituem ainda um ponto fraco da TRI (Muñiz, 1990; Hambleton et al., 1991). Entre eles, há o χ^2 e a análise dos resíduos. Hambleton, Swaminathan e Rogers (1991) acham a demonstração da adequação do modelo aos dados empíricos algo fundamental quando se trabalha com a TRI e se queixam das soluções puramente estatísticas que os autores vêm apresentando, soluções que, além de não serem satisfatórias, deixam de fora aspectos importantes a serem considerados nesta adequação. Eles afirmam que, pelo menos, três aspectos devem ser atendidos nesta problemática da adequação do modelo aos dados empíricos, a saber,

1) validade das suposições do modelo para os dados do teste;

2) extensão até a qual as propriedades esperadas do modelo (i.é, invariância dos parâmetros dos itens e da aptidão) são atingidas;

3) precisão das predições do modelo utilizando dados reais e, se for apropriado, simulações de testes.

Inclusive, eles apresentam uma série de possíveis técnicas para verificar os três aspectos mencionados, apresentadas na tabela 4-4.

Tabela 4-4. Enfoques para avaliar a adequação

Avaliando as Suposições do Modelo
1 – *Unidimensionalidade*
1) Plotar o *scree plot* dos *eigenvalues* para verificar a presença de um fator dominante (Reckcase, 1979)
2) Comparar o *scree plot* da matriz de intercorrelações dos itens do teste com o *scree plot* da matriz de intercorrelações de dados randômicos, sendo esta matriz composta de mesmo número de itens e sujeitos do teste original. Se houver unidimensionalidade, então os dois *scree plot* serão similares, mas o primeiro *eigenvalue* será bem maior no caso dos dados reais (Horn, 1965). Drasgow e Lissak (1983) trabalharam melhor esta técnica.
3) Analisar o pressuposto da independência local, verificando a matriz de variância-covariância ou a das correlações para sujeitos de diferentes intervalos na escala do teta ou dos escores do teste (McDonald, 1981; Tucker, Humphreys & Roznowski, 1986). Os elementos das matrizes fora da diagonal devem ser pequenos e próximos de zero se o pressuposto da unidimensionalidade for atingido.
4) Forçar uma análise fatorial não linear de um fator e verificar os resíduos (Hattie, 1985; McDonald, 1981).
5) Utilizar uma análise fatorial baseada na TRI (Bock, Gibbons & Muraki, 1988): trata-se de uma versão multidimensional do modelo de ogiva de três parâmetros e que explica o vetor das respostas aos itens.
6) Analisar num arquivo à parte os itens que parecem violar os pressupostos e em seguida analisá-los juntamente com o arquivo geral. Se os valores *b* destes itens em ambas as saídas foram lineares, então se pode supor que a unidimensionalidade está presente (Bejar, 1980).

2 – *Índices de Discriminação Iguais* (pressuposto do modelo de 1-parâmetro)

1) Se a correlação entre item e teste (bisserial ou ponto-bisserial) foi razoavelmente homogênea, então se pode presumir que os itens são igualmente discriminativos.

3 – *Chute Mínimo* (pressuposto do modelo de 2-parâmetros)

1) Verificar o desempenho de sujeitos de baixa aptidão nos itens difíceis: se seu desempenho for próximo de zero, pode-se supor que houve pouco chute.

2) Plotar os índices de regressão de item-teste (Baker, 1964, 1965): desempenho próximo de zero de sujeitos de escores baixos suportam o pressuposto de chute mínimo.

3) Verificar a dificuldade do teste, limites de tempo, formato do item para ver se tais fatores favorecem o chute.

4 – *Administração de Testes de Poder* (não testes de velocidade)

1) Verificar número de itens omitidos contra número de itens respondidos errados (Gulliksen, 1950): se a razão for próxima de zero, o teste é de poder.

2) Comparar os escores dos sujeitos respondendo o teste com tempo definido e tempo livre: se os escores se sobrepõem, então o teste é de poder.

3) Comparar as percentagens de completação do teste por 100%, 80% e 75% dos sujeitos: se quase todos os sujeitos completaram todos os itens, a velocidade será um fator irrelevante no teste.

Avaliando Características Esperadas do Modelo

Invariância das Estimativas dos Parâmetros da Aptidão

1) Comparar estimativas da aptidão para diferentes amostras de itens (por exemplo, itens fáceis e itens difíceis; itens cobrindo categorias diferentes do traço latente): se as estimativas não diferem muito, há invariância (Wright, 1968).

Invariância das Estimativas dos Parâmetros dos Itens

1) Comparar as estimativas dos parâmetros dos itens (a, b, c) para dois ou mais subgrupos da população alvo do teste (por exemplo, masculinos vs. femininos, diferentes níveis educacionais, diferentes regiões do país, etc.): se os plots desses parâmetros forem lineares, então há invariância. Pode-se verificar tal ocorrência, utilizando amostras randômicas equivalentes (Shepard, Camili & Williams, 1984).

Avaliando Predições do Modelo com Dados Reais e Dados Simulados

1) Investigar os resíduos da adequação do modelo aos dados: serve para escolher o melhor modelo TRI de análise (Hambleton & Swaminathan, 1985; Ludlow, 1985, 1986; Wright & Stone, 1979).

2) Comparar as distribuições dos escores observados e preditos pelo modelo: uma análise do χ^2 ou métodos gráficos mostram se as distribuições são similares (Hamphrey & Traub, 1973).

3) Investigar os efeitos da localização do item (Kingston & Dorans, 1984; Yen, 1980), efeitos de prática, rapidez e cola (Drasgow, Levine & Mehrens, 1987), má escolha do modelo (Wainer & Thissen, 1987), instrução recente (Cook, Eignor & Taft, 1988), variáveis de processamento cognitivo (Tatsuoka, 1987) e outros: poderão mostrar a adequação de uso de dado modelo da TRI.

4) Fazer um scatterplot das estimativas da aptidão e dos respectivos escores no teste: se a relação for forte (linear), a adequação do modelo aos dados é aceitável (Lord, 1974).

5) Aplicar miríades de testes estatísticos para determinar a adequação geral do modelo, dos itens, das pessoas (Andersen, 1973; Gustafsson, 1980; Ludlow, 1985, 1986; Traub & Wolfe, 1981; Wright & Stone, 1979, Yen, 1981; ...).

6) Comparar parâmetros estimados de itens e aptidão, utilizando métodos de simulação em computador (Hambleton & Cook, 1983).

7) Investigar a robustez do modelo com métodos de simulação em computador (Ansley & Forsyth, 1985; Drasgow & Parsons, 1983).

Você vê que os métodos para verificar a adequação do modelo da TRI aos dados são sem fim. Somente para verificar a unidimensionalidade, por exemplo, Hattie (1984, 1985) apresenta 87 índices diferentes. É uma selva e cada autor tem suas preferências na escolha das técnicas de adequação. De qualquer forma, o uso prático dessas técnicas exige o computador e programas específicos; quando você for utilizá-los, veja qual das técnicas está sendo utilizada e para verificar que tipo de adequação: caso seja a unidimensionalidade, a invariância das estimativas dos parâmetros, etc. Para dar a você um pouco de gosto destas técnicas, veja as duas que em seguida explico.

5.1 – O qui quadrado (χ^2)

Wright e Panchapakesan (1969, in: Muñiz, 1990) utiliza uma estatística parecida com o χ^2 para verificar o ajuste do modelo, cuja fórmula é

$$\chi^2 = \sum_{j=1}^{k} \frac{n_i [P(\theta_i) - Pe(\theta_j)]^2}{[P(\theta_i)][1 - P(\theta_j)]} \qquad (4.34)$$

onde

k = número de categorias em que se dividiu o θ

n_i = número de sujeitos dentro da categoria

$P(\theta_j)$ = valor da CCI para a categoria *j* dado pela fórmula do modelo

$Pe(\theta_j)$ = proporção de sujeitos que de fato acertaram o item para a categoria *j*.

χ^2 distribui-se com k-1 graus de liberdade.

Exemplo (Muñiz, 1990: 51-53): Mil sujeitos responderam 20 itens. O programa LOGIST estimou os parâmetros, mostrando que o modelo de dois parâmetros seria o aconselhável. Para o item 10, o programa deu que a = 1 e b = 2. O θ foi dividido em 5 categorias (usa-se o ponto médio das categorias para os cálculos) e os resultados foram os da tabela 4-5.

Tabela 4-5. Proporção de 1.000 sujeitos acertando item 10 por categoria de θ

θ	n_i	$Pe(\theta_j)$	$P(\theta_j)$
4-5	70	0,97	0,99
3-4	90	0,95	0,92
2-3	200	0,70	0,70
1-2	300	0,35	0,30
0-1	340	0,10	0,07
	1.000		

Os valores P(θ) calculados pelo modelo de dois parâmetros [$P(\theta_j)$] foram conseguidos usando a fórmula deste modelo, onde os parâmetros para o item 10 foram: a = 1 e b = 2.

Aplicando-se a fórmula do χ^2 resulta:

$$\chi^2 = \frac{340(0,07-0,10)^2}{(0,07)(1-0,07)} + \frac{300(0,30-0,35)^2}{(0,30)(1-0,30)} + \frac{200(0,70-0,70)^2}{(0,70)(1-0,70)} + \frac{90(0,92-0,95)^2}{(0,92)(1-0,92)} + \frac{70(0,99-0,970^2}{(0,99)(1-0,99)}$$

$$= 4,70 + 3,57 + 0,00 + 1,10 + 2,83$$

$$= 12,2$$

que, para graus de liberdade k-1 = 5-1 = 4, a probabilidade de tal χ^2 (= 12,2) a gl = 4 ocorrer por acaso se situa entre 0,02 e 0,01. Portanto, somente ao nível de 98% de confiança pode-se afirmar que o modelo de dois parâmetros se ajusta aos dados empíricos do item 10. Pela figura 4-8, vê-se que os valores empíricos, os P(θ), e os calculados ou esperados, Pe(θ), são bastante similares.

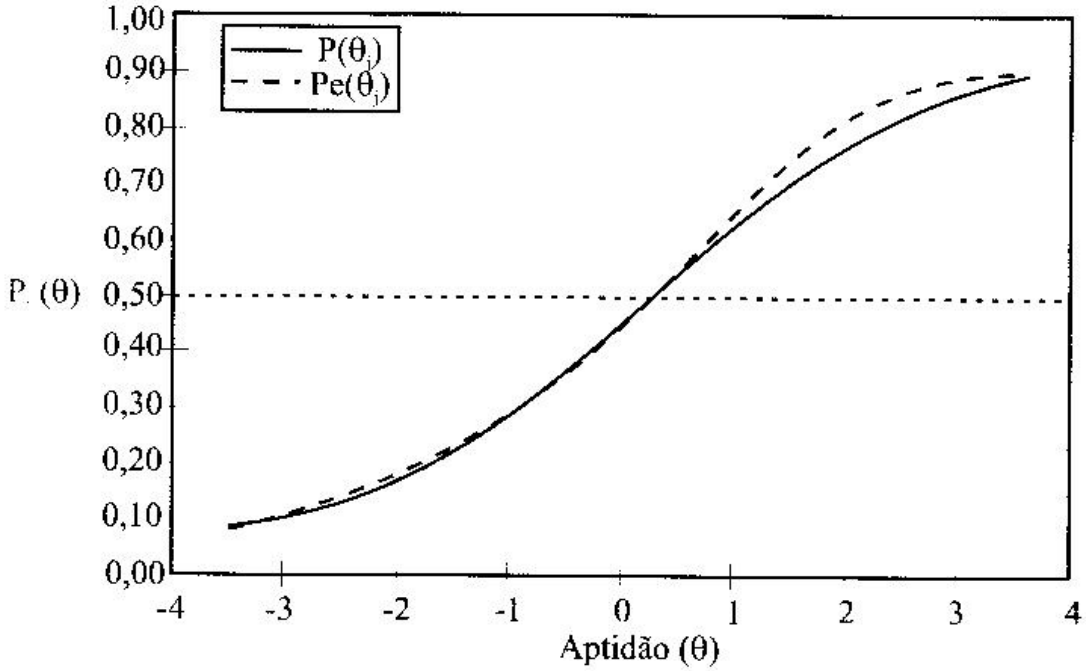


Figura 4-8. Valores preditos pelo modelo $Pe(\theta_j)$ e valores empíricos $P(\theta_j)$

5.2 – Análise dos resíduos

Esta análise consiste em verificar se a diferença entre o desempenho real dos sujeitos num item e o desempenho predito pelo modelo não é estatisticamente diferente de 0; sua fórmula é

$$r_{ij} = P_{ij} - E(P_{ij}) \qquad (4.35)$$

onde

r_{ij} = o resíduo

P_{ij} = desempenho real, isto é, proporção de respostas corretas ao item *i* na categoria *j* da aptidão θ.

$E(P_{ij})$ = desempenho predito pelo modelo.

Normalmente, este resíduo é expresso em dados padronizados (resíduo padronizado = z_{ij}), sendo sua fórmula a seguinte:

$$z_{ij} = \frac{P_{ij} - E(P_{ij})}{\sqrt{E(P_{ij})[1 - E(P_{ij})]/N_j}} \qquad (4.36)$$

onde N_j é o número de sujeitos na categoria. Esta categoria se refere a que o θ deve ser dividido em categorias (10 a 15) como no caso do χ^2.

A figura 4-9 mostra que os dados empíricos não se coadunam com os preditos, pois as duas linhas não coincidem; de fato os dados preditos se afastam de um desvio-padrão com respeito aos dados empíricos.

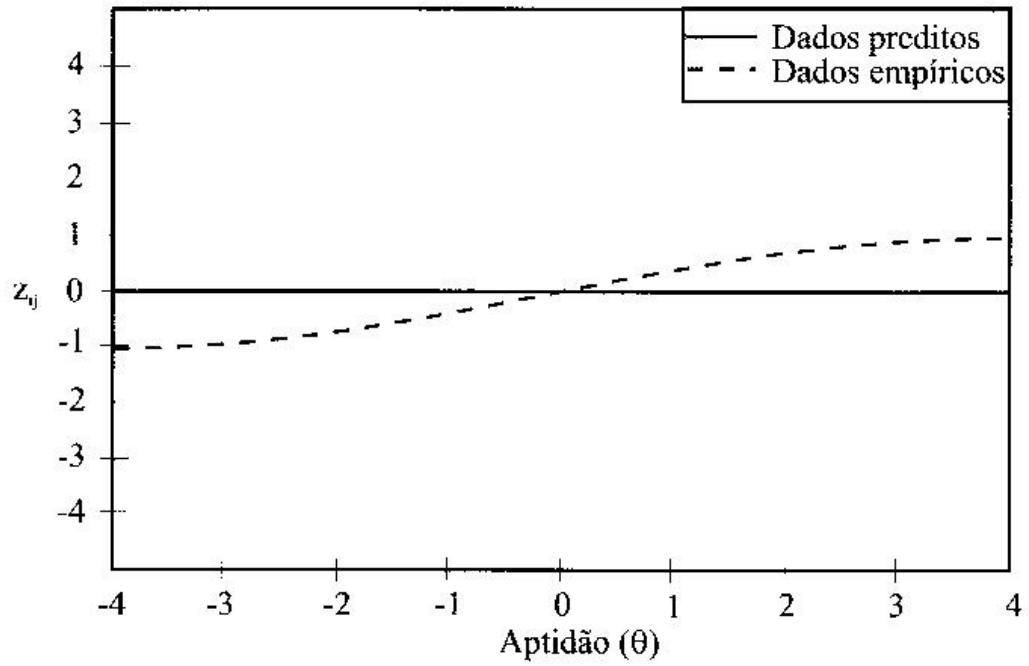


Figura 4-9. Discrepância entre modelo e dados empíricos

6 – Invariância dos parâmetros

A invariância dos parâmetros constitui o ponto central da TRI e afirma que se pode estimar: (1) os escores dos sujeitos independentemente do teste utilizado e (2) os parâmetros dos itens independentemente da amostra de sujeitos utilizada. Se o modelo TRI utilizado se adequa aos dados empíricos, então são resguardados estes objetivos, ilustrados na figura 4-10.

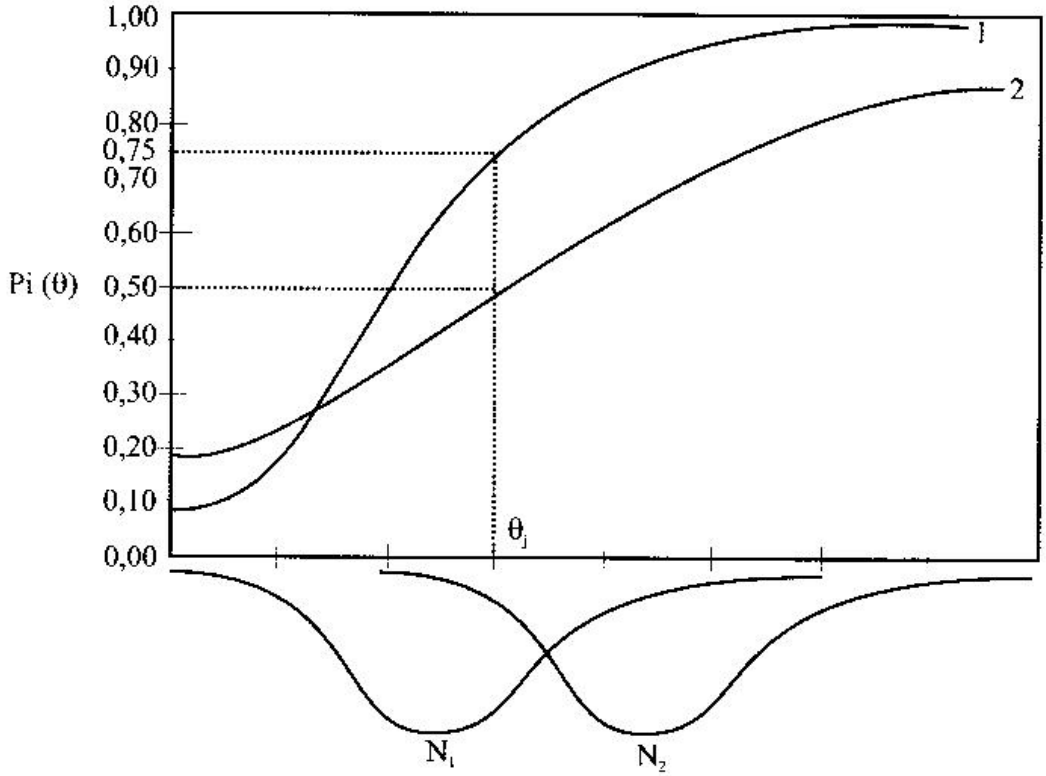


Figura 4-10. CCI para dois itens e distribuição de θ para dois grupos de sujeito

A figura 4-10 mostra que

1) as curvas CCI para os dois itens podem ser obtidas tanto com a amostra N_1 quanto a N_2; consequentemente, os parâmetros dos itens independem da amostra utilizada e

2) o valor da aptidão θ_j pode ser obtido utilizando-se tanto o item 1 quanto o item 2; este θ_j corresponde à probabilidade de acerto de 75% do item 1 e 50% do item 2; portanto, o θ independe dos itens utilizados.

Continua valendo, contudo, como em qualquer estimação estatística, que quanto maior e mais heterogênea a amostra de sujeitos, mais precisa será a estimação dos parâmetros.

Para demonstrar a invariância da aptidão (θ), aplicam-se dois testes com itens diferentes, mas que medem a mesma aptidão, a uma mesma amostra de sujeitos e os resultados mostrarão se há ou não coincidência. Se houver coincidência, então os itens dos dois testes se distribuirão em torno de uma linha reta num sistema de coordenadas como na figura 4-11 e uma indicação numérica será dada pela correlação de Pearson entre as duas avaliações.

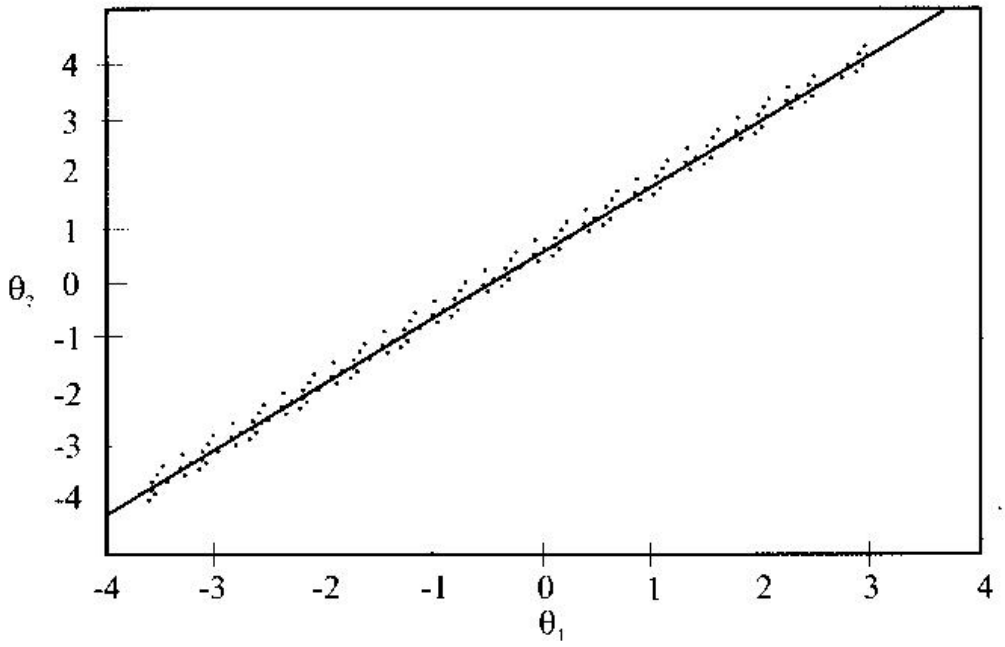


Figura 4-11. Valores θ obtidos por testes diferentes

Evento similar ocorre na demonstração da invariância dos parâmetros dos itens (a, b, c). Aqui se usam duas amostras de sujeitos para responder ao mesmo teste. Se os parâmetros dos itens são os mesmos nas duas amostras, novamente surgirá uma linha reta nas coordenadas (agora expressas pela amostra 1 e pela amostra 2 de sujeitos, em lugar de θ_1 e θ_2 da figura 4-11) e a correlação de Pearson estima a coincidência dos parâmetros.

7 – Aplicações da TRI[11]

Entre as inúmeras possíveis aplicações da TRI na teoria dos testes, algumas são especialmente relevantes, nas quais a TRI tem contribuições inovadoras e promissoras (delas falaremos mais no cap. 10). Entre elas, vamos salientar as seguintes:

7.1 – Banco de itens

Na era do computador, a existência de banco de itens permite uma utilização, não somente mais sofisticada, mas muito mais expedita, prática e eficiente da medida psicológica. A construção do banco de itens é viável dentro da teoria psicométrica clássica, fornecendo os parâmetros de dificul-

---

11. Sobre estas aplicações cf. detalhes no capítulo 10.

dade e discriminação de cada item. Entretanto, nesta teoria, os parâmetros são dependentes da amostra de sujeitos utilizados. A TRI permite estabelecer os mesmos parâmetros independentemente da amostra utilizada; daí é possível incluir sempre novos itens diretamente comparáveis aos já inclusos no banco. A técnica para esta façanha, entre outras, consiste em aplicar os novos itens juntamente com uma amostra de itens já incluídos no banco a uma amostra razoável de sujeitos e estimar os parâmetros dos novos itens em confronto com os dos itens utilizados do banco de itens. Assim os novos itens entram no banco nas mesmas condições que os velhos.

7.2 – Testes sob medida (computerized adaptive testing – CAT)

Um teste fornece uma medida mais precisa da aptidão (θ) do sujeito quando seus itens se situam no nível de dificuldade correspondente ao nível da aptidão do sujeito. Assim, um sujeito com $\theta = 2$ deveria ser examinado com itens de dificuldade (b_i) em torno de 2, dado que a função de informação deles[12], para tal θ, é máxima a esse nível. Na utilização tradicional dos testes, é costumeiro aplicar-se um mesmo teste a sujeitos de níveis diferentes de θ. Num caso destes, o teste avalia bem alguns sujeitos e mal outros. O ideal seria aplicar para cada sujeito um teste diferente, obviamente medindo a mesma aptidão, mas que se emparelhe, em termos de dificuldade, ao nível θ de cada sujeito. Esta é a ideia atrás dos testes feitos sob medida (*tailored* ou *computer adaptive testing*). Nesta situação, a sequência dos itens submetidos ao examinando depende do desempenho do mesmo no item anterior; assim, para cada sujeito, a sequência de itens é diferente. O que fica como problema a ser resolvido em tal procedimento é garantir que sujeitos diferentes submetidos a itens diferentes estejam sendo avaliados da mesma forma, isto é, que os itens estejam medindo o mesmo traço latente para todos os sujeitos. A TRI, podendo estabelecer os parâmetros e as funções de informação dos itens, pode também demonstrar a equivalência entre testes diferentes, ou melhor, selecionar, para sujeitos diferentes, itens diferentes, mas equivalentes.

12. A função de informação é tratada no capítulo 6 (Validade dos testes psicológicos).

CAPÍTULO 5
Análise dos itens

Introdução

Os itens constituem a representação comportamental do traço latente. Eles são as tarefas, ações empíricas através das quais o traço latente se manifesta. Na verdade, como o comportamento (verbal, motor) é o único nível em que se pode trabalhar cientificamente (empiricamente) em Psicologia, é neste nível que se deve procurar a solução para o problema da representação e, portanto, do conhecimento dos processos latentes.

Como o comportamento representa estes traços latentes? É o problema das definições operacionais. A Psicometria responde a esta questão pela análise de uma série de parâmetros que os comportamentos (tipicamente chamados de itens) devem apresentar, os quais foram brevemente apresentados no capítulo 3 e que serão aqui desenvolvidos.

Na verdade, há dois tipos de análise de itens, que poderíamos chamar de análise teórica e análise empírica ou estatística, sendo que esta última pode ser expressa em termos geométricos ou algébricos, isto é, há uma análise gráfica dos itens e há uma análise algébrica. A análise gráfica se baseia na TCT ou no escore total e análise algébrica, além da TCT, tem a análise da TRI, como veremos.

I – A ANÁLISE TEÓRICA DOS ITENS

Esta análise é feita por juízes e visa estabelecer a compreensão dos itens (análise semântica) e a pertinência dos mesmos ao atributo que pretendem medir. Esta última é, às vezes, chamada de análise de conteúdo, mas propriamente deve ser chamada de análise de construto, dado que precisamente procura verificar a adequação (conformidade) da representação comportamental do(s) atributo(s) latente(s).

1.1 – Análise semântica dos itens

No caso da *análise semântica*, os juízes são sujeitos da própria população para a qual se quer construir o teste. Duas preocupações são relevantes nesta análise: (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; e (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). 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.

Como realizar esta análise? Uma das maneiras mais eficazes para testar a compreensão das tarefas (itens) consiste em verificá-las numa situação de entrevista com pequenos grupos (3 a 4) em atmosfera de *brainstorming*. O experimentador apresenta os itens um a um e pede ao grupo para reproduzi-los com as próprias palavras. Se a reprodução não for consenso ou não corresponder ao que o experimentador pretendia com o item, este obviamente tem problemas de compreensão e deve ser reformulado ou abandonado. Nesta mesma sessão surge, na maioria das vezes, a formulação correta da tarefa oferecida pelo próprio grupo. Itens que com dois grupos de sujeitos não apresentam problemas de compreensão não precisam mais ser ulteriormente checados. Itens com problemas nesta área poderão exigir até meia dúzia de conferências, depois disso o item que ainda persistir duvidoso deve ser descartado.

1.2 – Análise dos juízes

Na *análise do conteúdo do teste*, os juízes 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. A tarefa deles consiste em dizer se o item constitui uma representação adequada de tal ou tal fator (traço latente). Por exemplo: construiu-se um teste com 30 itens para cobrir três traços latentes (afiliação, ansiedade, autonomia), eu tenho a minha versão de qual fator cada item está representando. Preciso da opinião de outros especialistas para confirmar ou não minha versão. Assim, peço a uma meia dúzia

deles para individualmente opinarem a qual dos fatores cada item se refere. Para tanto, dou aos especialistas duas listas: uma com as definições do que eu entendo por cada um dos fatores e outra lista, em forma de tabela de dupla entrada, com o elenco dos itens alistados à esquerda e os fatores no topo. A tarefa dos especialistas consistirá em lançarem uma marca para cada item sob o fator do qual o item se constitui representante, conforme tabela 5-1.

Tabela 5-1. Listagem de itens e fatores para análise de juízes

	Fatores		
Itens	Afiliação	Ansiedade	Autonomia
1 – Gosto de mandar			X
2 – Tenho medo do escuro		X	
3 – Gosto de amigos	X		
....	...	...	...
30 – Detesto ver sangue		X	

A análise destas listas consiste em verificar se há concordância entre os juízes. O item é retido no elenco se houver uma concordância de cerca de 80% entre os juízes. Assim, se o item 1 foi assinalado por 8 entre 10 juízes como sendo representante do fator autonomia, então ele conseguiu concordância (8 / 10 = 0,80) e é retido no elenco.

II – A Análise Empírica dos Itens

Introdução

A análise empírica dos itens implica na avaliação de uma série de parâmetros que eles devem possuir a fim de se apresentarem como tarefas adequadas para o que o teste se propõe medir. Estes parâmetros podem ser elencados nos seguintes: unidimensionalidade, dificuldade, discriminação, vieses (entre estes, particularmente o chute e a função diferencial, isto é, o DIF), tendenciosidade de resposta, validade, precisão.

Dessas características dos itens, a Psicometria tradicionalmente analisa a dificuldade, a discriminação e o chute. Com a globalização e consequente uso dos testes entre diferentes culturas, a Psicometria vem se preocupando mais e mais com o estudo dos vieses dos itens devidos a fatores de ordem cultural e sociocultural, tema que vem caracterizado como a *função diferencial dos itens (differential item functioning* – DIF). Estes parâmetros dos itens são concebidos e tratados de forma diferente pela Teoria Clássica dos Testes (TCT) e pela Teoria de Resposta ao Item (TRI), como veremos.

A análise dos parâmetros dos itens se faz em cima dos dados coletados de uma amostra de sujeitos representativa da população para a qual o teste está sendo ou foi construído, utilizando-se análises estatísticas apropriadas. Há uma série de técnicas de análise para efetuar a avaliação de todos esses parâmetros dos itens. Essas análises vão desde uma pura análise geométrica (análise gráfica dos itens) até as técnicas da análise fatorial, da TCT, da TRI e das análises do DIF. Para organizar essas análises, vamos tratar o tema sob dois ângulos, a saber: (1) a análise gráfica ou geométrica dos itens e (2) as análises algébricas dos itens, entre as quais entra a análise fatorial e as análises ditadas pelos modelos da TCT e da TRI. Neste particular, a TRI fez progressos notáveis na análise dos itens, tornando bastante obsoletas as análises que a TCT fazia, mas ambos os tipos de análises serão apresentados, dado que muitos testes no mercado ainda hoje em dia apresentam dados de análises feitas através da TCT e, mesmo, porque estas análises substituem as da TRI quando este modelo não for adequado aos dados empíricos (na falta de *data-model goodness-of-fit*).

A exposição sobre esses parâmetros dos itens usará como exemplo o teste TNVRA[1].

1. O TNVRA (Teste Não Verbal de Raciocínio para Adultos) consta de 30 itens para medir o raciocínio dedutivo de adolescentes e adultos. Ele foi construído a partir das Matrizes Progressivas de Raven adulto. Está presentemente sendo validado no Laboratório de Pesquisa em Avaliação e Medida – LabPAM, da UnB.

A – A Análise Gráfica dos itens

Jacob A. Laros, Ph.D., UnB

A análise gráfica dos itens utiliza o modelo da TCT e é especialmente utilizada com testes de aptidão. Ela se baseia em dois pressupostos: (1) um sujeito que dá resposta certa a um item de múltipla escolha sabe mais, em geral, que um sujeito que dá uma resposta errada e (2) um sujeito que acerta mais itens sabe mais do que um sujeito que acerta menos itens.

Desses pressupostos segue que o sujeito que sabe mais terá um escore total no teste maior que um sujeito que sabe menos. Assim, pode-se inferir que sujeitos com escores maiores tenderão a acertar mais um certo item que sujeitos com escores menores. Sendo isto verdade, então é possível analisar os itens de um teste em função do escore total no teste, verificando se, com o aumento do escore total, também aumenta a probabilidade de resposta correta de um dado item e, conversamente, diminui a resposta errada. Ocorrendo tal evento, então se pode afirmar que o item é de boa qualidade. Esta é a lógica da análise gráfica dos itens. A figura 5-1 mostra um exemplo teórico de um item de boa qualidade.

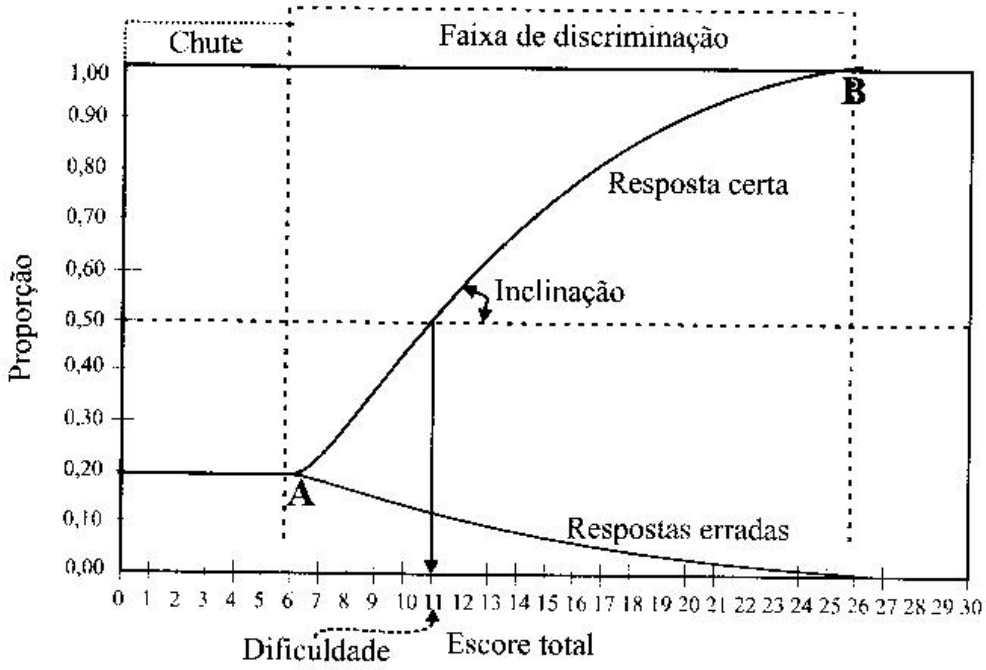


Figura 5-1. A curva gráfica do item

A figura mostra, na abscissa, o escore total dos sujeitos num teste de 30 itens (este escore vai de 0 a 30) e, na ordenada, mostra a proporção de sujeitos que deram a resposta certa e as respostas erradas a um dado

item do teste. Você vê que, idealmente, com o aumento do escore total, a resposta certa ao item deve aumentar de um modo sistemático, enquanto as respostas erradas devem diminuir também sistematicamente. As alternativas de resposta para os itens deste teste eram de 5; de sorte que cada uma delas tem a chance de ser escolhida aleatoriamente em 20% das vezes. Desta forma, se os sujeitos respondessem de um modo aleatório ao teste, eles obteriam um escore total de 6; pode-se, assim, dizer que até a um escore total de 6, todas as 5 alternativas de resposta de um dado item têm o mesmo nível de ocorrência, a mesma chance de serem escolhidas, isto é, 20%. Esta faixa do escore total, que vai de 0 a 6, é a faixa chamada de *chute*, porque qualquer um desses escores totais poderia ser obtido simplesmente chutando as respostas aos itens do teste. É a partir deste escore total 6 que vai se verificar o aumento sistemático da resposta certa e a diminuição das alternativas incorretas, se o teste não foi respondido de uma forma aleatória pelos sujeitos. A resposta certa, assim, vai aumentando até atingir o platô de p = 1,00, que no nosso caso aconteceu com o escore total 26. Isto é, do escore 26 em diante já não há mais ganhos na resposta certa, porque já atingiu o 100% de ocorrência.

Os escores totais de 6 e 26, no caso do item da figura 5-1, são dois pontos importantes, porque eles definem a faixa dentro da qual o item é capaz de discriminar sujeitos com escores diferentes: é chamada a faixa de *discriminação* do item. A faixa começa onde a curva da resposta certa já não é mais ultrapassada por alguma das curvas que expressam as respostas erradas e acaba quando atingiu pela primeira vez a frequência máxima de 100% de acertos (estes dois pontos estão marcados por **A** e **B** na figura 5-1). Fora desta faixa, isto é, abaixo de 6 ou acima de 26, o item não consegue discriminar diferenças entre os escores dos sujeitos. Agora, quanto mais rápido sobe a linha da resposta certa, mais discriminativo é dito ser o item; e tal fato é definido pelo ângulo de incidência da linha ao atravessar a probabilidade de 50% de chance de esta resposta certa ser dada. Este ângulo é dito a *inclinação* da curva.

Este ponto em que a curva do item atravessa a linha que representa a chance de 50% da resposta certa ser dada representa a *dificuldade* geral do item no teste. Se você baixar uma perpendicular deste ponto até a abscissa, você terá o escore total que define a dificuldade do item; no nosso caso, ela é de 11 num teste que pode dar um escore total de 30. Agora, para saber se um tal escore de 11 de dificuldade diz ser o item fácil ou difícil, é preciso saber se o teste todo é fácil ou difícil, isto é, qual é a média

geral do teste. Mesmo assim, você pode saber a dificuldade relativa de todos os itens do teste e pode ordená-los em termos de dificuldade dentro do teste, conhecendo esta dificuldade individual de cada item.

Concluindo: A análise gráfica dos itens dá três informações importantes sobre cada um deles, a saber:

- dificuldade: perpendicular que parte do ponto de 50% de ocorrência da resposta certa sobre a abscissa;
- discriminação: ângulo de incidência da curva da resposta certa no ponto de 50% de ocorrência desta resposta;
- chute: escore total que corresponde à percentagem de ocorrência de respostas aleatórias, dependendo do número de alternativas de respostas para o item. No caso de 30 itens, com 5 alternativas de respostas, o chute corresponde a um escore total de 6.

As figuras 5-2 a 5-5 mostram as configurações gráficas de vários tipos de itens: item de boa qualidade (figura 5-2), item de má qualidade (figura 5-3), item fácil (figura 5-4) e item difícil (figura 5-5). A resposta correta para os itens é a linha de cor vermelha; as linhas pontilhadas indicam a faixa de discriminação do item e a linha assinalada com seta indica a dificuldade do mesmo.

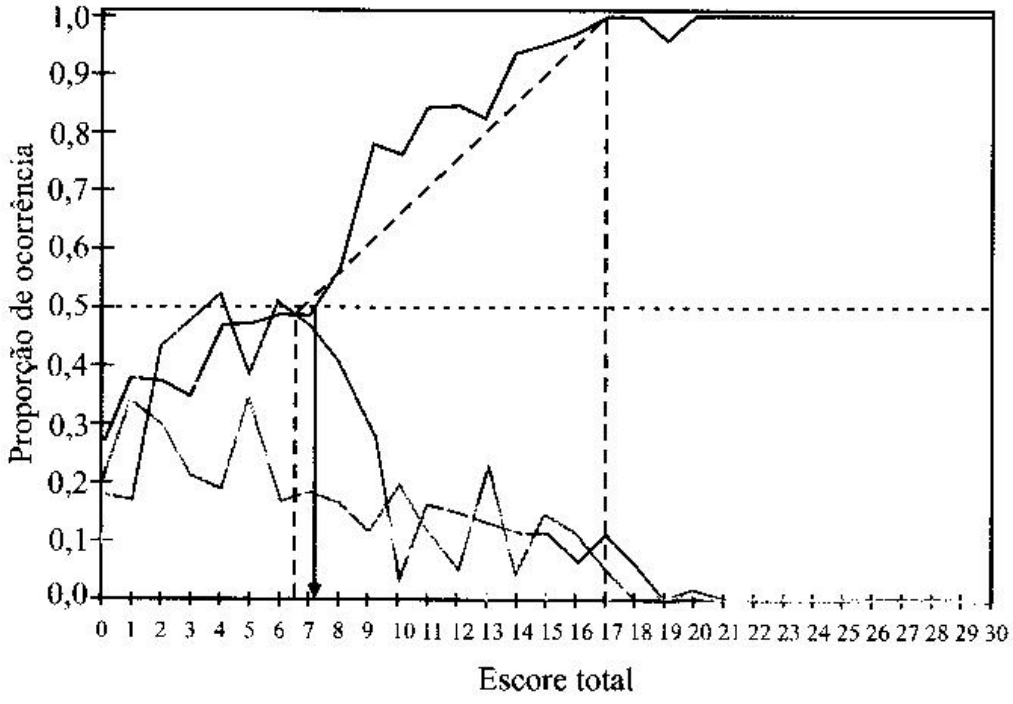


Figura 5-2. Análise gráfica de um item de boa qualidade

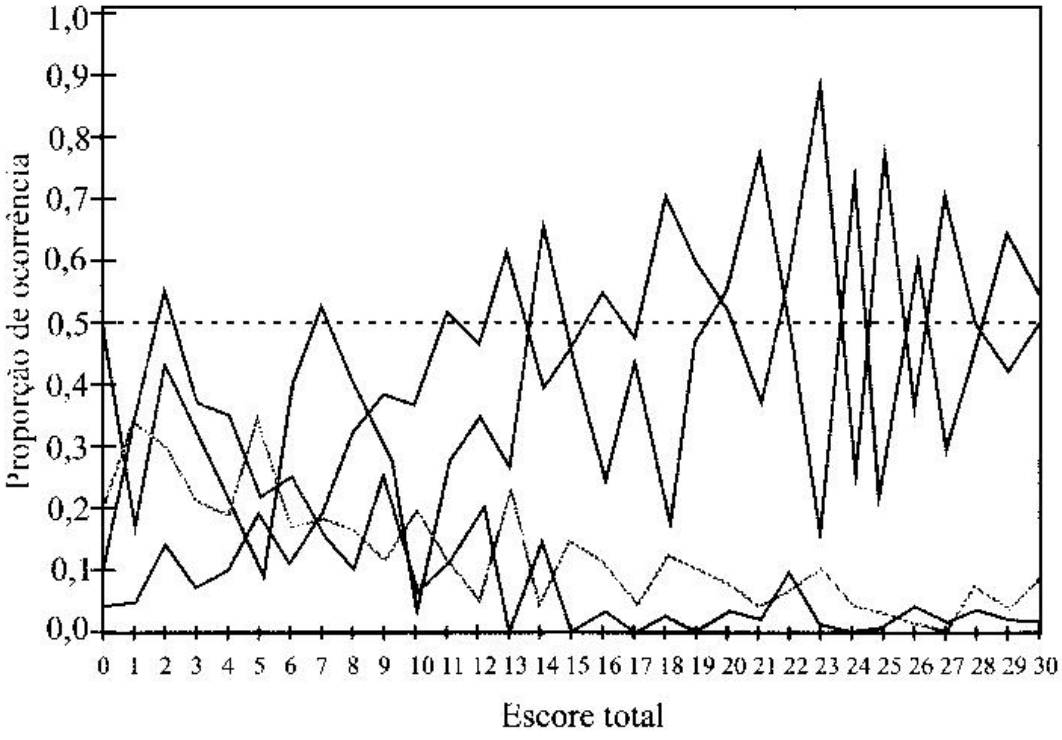


Figura 5-3. Análise gráfica de um item de má qualidade

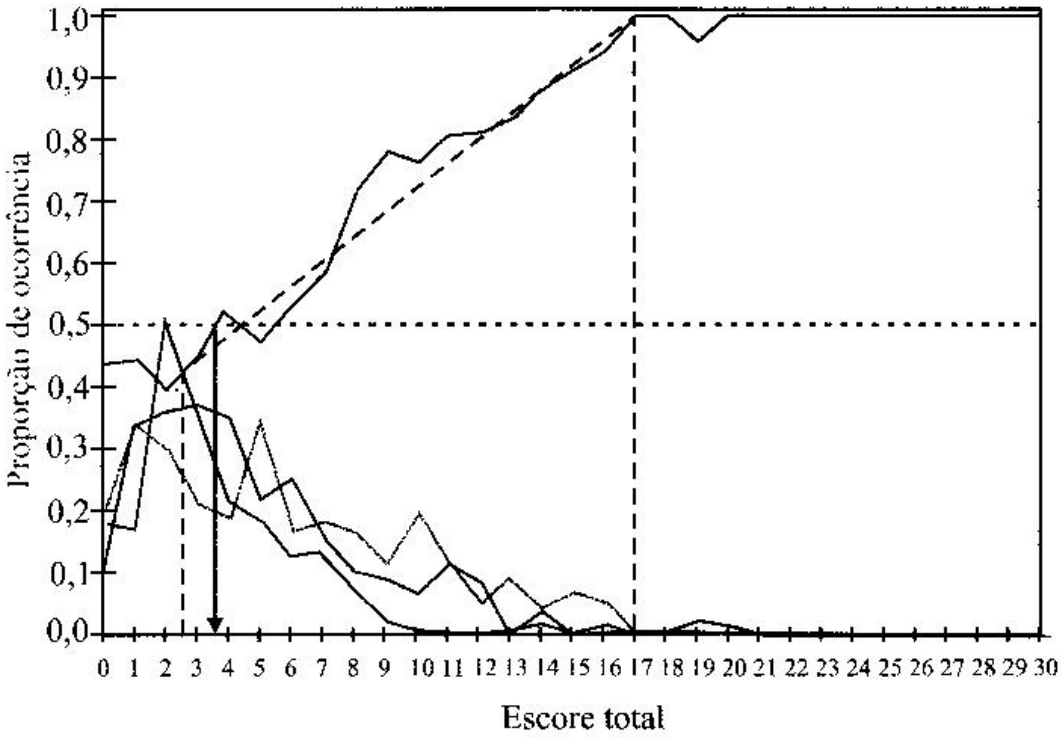


Figura 5-4. Análise gráfica de um item muito fácil

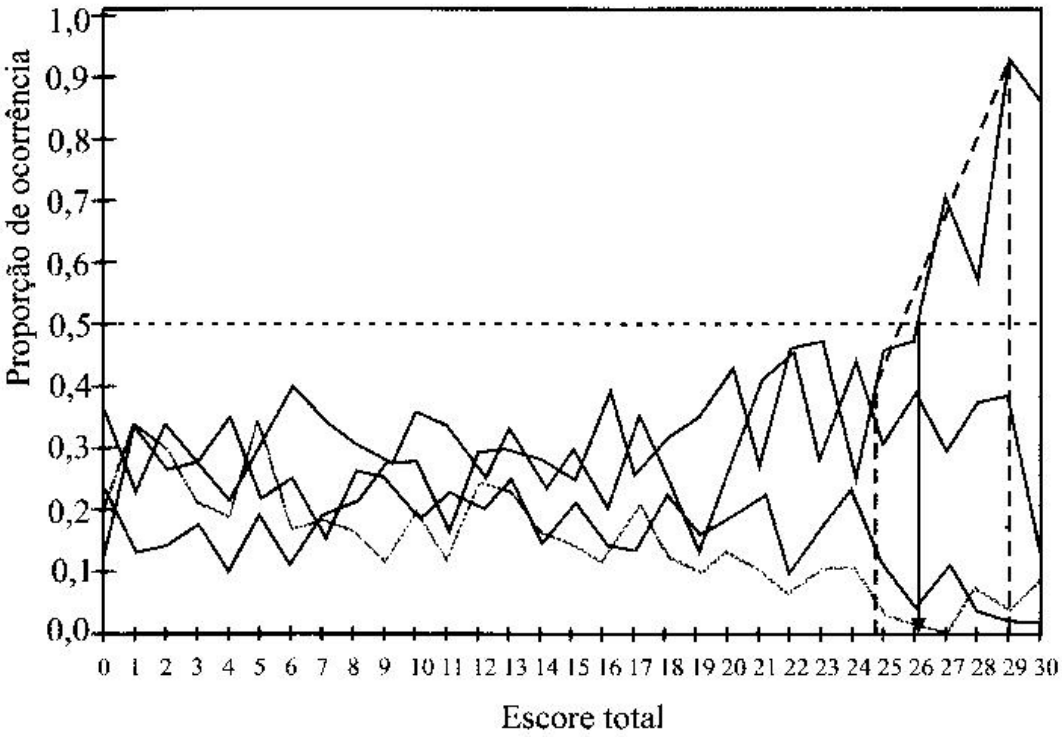


Figura 5-5. Análise gráfica de um item muito difícil

B – A Análise Algébrica dos Itens

A análise algébrica dos itens faz uso de algoritmos estatísticos para avaliar vários aspectos dos itens, em particular, a unidimensionalidade, a dificuldade, a discriminação, os vieses, a validade e a precisão. Novamente, as análises neste setor são variadas e às vezes bastante complexas de um ponto de vista estatístico; ademais, você vai encontrar, inclusive, divergências entre os peritos quanto à adequação de algumas dessas análises. Mas vamos lá, tratando desse tema pela discussão de cada um dos parâmetros dos itens.

B.1 – Unidimensionalidade dos Itens[2]

Quando se está analisando uma série de itens, tanto a TCT quanto a TRI supõem que todos eles estejam medindo a mesma coisa. No caso da TCT, isto ocorre porque ela trabalha com o escore total e cada item é avaliado em função deste escore total; acontece, porém, que o escore total consiste na soma das respostas dadas aos itens; assim, ela faz a suposição que eles são somáveis e isto faz sentido somente se eles se referem à mesma coisa, pois não dá para somar alhos e bugalhos. Alhos e bugalhos

2. Para uma discussão mais detalhada deste tema, cf. Laros, Pasquali e Rodrigues(2000).

dão um agregado, não uma soma. No caso da TRI, todos os itens são avaliados em função de um traço latente; assim, todos eles devem se referir a este traço latente.

Este é o problema chamado de unidimensionalidade, isto é, os itens estarem medindo uma única e a mesma coisa. Você vê, deste arrazoado, que a unidimensionalidade dos itens de um teste é uma condição necessária para analisar qualquer característica ulterior de um item, tais como a dificuldade e discriminação do mesmo.

Como a unidimensionalidade é algo central na análise dos itens de um teste, você já pode imaginar que os pesquisadores da área inventaram mil maneiras de resolver o problema. Mas, infelizmente, não existe ainda um índice efetivo e aceito por todos para a solução desse problema, apesar da necessidade urgente de tais índices (Hambleton, Swaminathan, Cook, Eignor & Gifford, 1978; Jones, Sabera & Trosset, 1987; Lord, 1980; Hattie, 1985). De fato, Hattie (1985, apresentado em Laros, Pasquali & Rodrigues, 2000) identificou mais de 30 índices diferentes utilizados para abordar a questão da unidimensionalidade de uma série de itens. Ele categorizou esses índices dentro de cinco grupos, a saber: (1) índices baseados em padrões de resposta (Lumsden, 1959); (2) índices baseados na fidedignidade (Cortina, 1993; McDonald, 1981; Green, Lissitz & Mulaik, 1977); (3) índices baseados na análise de componentes principais (Bejar, 1980; Cook & Eignor, 1984; Cook, Dorans & Eignor, 1988; Hambleton & Rovinelli, 1986; Hulin, Drasgow & Parsons, 1983; Jones et al., 1987; McDonald & Ahlawat, 1974; Zwick, 1985); (4) índices baseados na análise fatorial e (5) índices baseados na TRI. Como existem muitas críticas com respeito aos métodos 1 a 3, vamos aqui expor apenas os métodos que dizem respeito à análise fatorial e à TRI. Em seguida será exposto um novo método que faz uso de princípios da análise fatorial e da TRI, chamado de análise fatorial *full information*, o qual parece atualmente como o mais recomendável no contexto da análise da unidimensionalidade de uma série de itens.

1 – Unidimensionalidade baseada na análise fatorial

Para a compreensão aprofundada da análise fatorial é preciso consultar livros especializados, pois não é possível expor tal técnica neste livro; contudo, vamos fazer uma breve exposição para entender a forma e a utilidade da mesma no contexto da análise da dimensionalidade de um conjunto (matriz) de dados. Para aprofundamento do método cf. Pasquali e Harman (1957).

A análise fatorial tradicional, abreviada como AF, consiste numa série de técnicas estatísticas que trabalha com análises multivariadas e matrizes. A matriz que a AF trabalha é tipicamente uma matriz de intercorrelações entre uma série de variáveis ou itens. A análise que ela faz consiste em verificar se uma série de itens pode ser reduzida idealmente a uma única dimensão ou variável, que ela chama de fator, com o qual todas as variáveis da série estão relacionadas. Sendo este o caso, então se conclui que os itens são unidimensionais, isto é, estão medindo a mesma coisa, que é o que o princípio da unidimensionalidade procura. A relação que cada item tem com o fator é expressa através da covariância ou da correlação; esta relação se chama de carga fatorial. Itens da série que têm alta carga no fator são itens unidimensionais, pois medem o mesmo fator, enquanto itens com cargas perto de 0 são itens estranhos e, por isso, devem ser descartados, porque não estão medindo a mesma coisa que os demais; estes itens pecam contra a unidimensionalidade e, portanto, não podem ser analisados juntamente com os outros. Cf. o exemplo da tabela 5-2.

Tabela 5-2. Matriz fatorial do TNVRA

Item	Carga
1	0,44
2	0,75
3	0,79
4	0,71
5	0,44
6	0,79
7	0,80
8	0,73
9	0,85
10	0,80
11	0,72
12	0,82
13	0,77
14	0,85
15	0,78
16	0,67
17	0,71
18	0,82
19	0,87
20	0,69
21	0,83
22	0,74
23	0,89
24	0,29
25	0,43
26	0,65
27	0,66
28	0,64
29	0,36
30	0,57

Nesta tabela se vê que a maioria dos itens do TNVRA está se referindo a uma única coisa, no caso, ao raciocínio dedutivo que este teste pretende medir. Isto porque os itens têm carga alta no fator. Carga alta significa um valor de pelo menos 0,30 e, no caso de testes de aptidão, estas cargas devem todas ter o mesmo sinal (no nosso caso, sinal positivo). Você vê, então, que apenas o item 24 não satisfaz a condição de carga alta e, por isso, não estaria medindo a mesma coisa que os demais. Isto significa que ele não se entende com os demais e, portanto, não faz parte do teste, devendo ser descartado; isto é, ele não satisfaz o critério da unidimensionalidade. Assim, se você for analisar as outras características dos itens do TNVRA, você deverá desconsiderar o item 24, porque ele não pode ser somado aos demais (se você for trabalhar com a

TCT) e não está medindo o mesmo traço latente que os demais (se for trabalhar com a TRI), embora ele esteja muito próximo de ser um item sofrivelmente adequado.

Entretanto, os autores, em geral, encontram alguns problemas com a análise fatorial tradicional (bem como com a análise dos componentes) para decidir a questão da unidimensionalidade dos itens. A primeira das preocupações consiste em que estes métodos trabalham com equações lineares, supondo uma relação linear entre as variáveis. Se a relação, contudo, for não linear, tanto a análise fatorial quanto a análise dos componentes principais produzem resultados estranhos que não podem ser interpretados ou que são simplesmente errôneos. Uma saída para este problema seria trabalhar com a análise fatorial não linear, mas a literatura é pelo menos ambígua sobre a eficácia desse método (Hattie, 1985). Um segundo problema surge quando se utilizam estas técnicas para analisar itens dicotômicos, isto é, itens que somente têm dois valores, tais como certo e errado, como é o caso com os testes de aptidão e desempenho em treinamento. Isto porque, neste caso, a matriz das intercorrelações entre os itens que vai ser analisada é constituída de correlações phi ou tetracóricas. Só que a correlação phi supõe que as variáveis sejam realmente dicotômicas e a tetracórica que elas tenham uma distribuição normal bivariada. São muitas suposições para assumi-las tranquilamente numa análise de itens. Mas, se tais suposições não forem satisfeitas, então as análises irão produzir resultados estranhos e não confiáveis. Além disso, há um terceiro problema, a saber, como decidir se uma dada matriz tem um ou mais fatores? Novamente entre os autores parece que cada qual tem sua opinião e inventam índices e mais índices para decidir a questão.

Deixando de lado esses melindres e especiosidades dos teóricos e estatísticos quanto à adequação da análise fatorial para decidir a questão da unidimensionalidade, se você a for utilizar para tal fim e supõe que a matriz é unifatorial, então peça a extração de um fator e veja se a grande maioria dos itens tem carga alta no fator. Se, contudo, grande parte dos itens não tem carga alta no fator, então siga extraindo mais fatores até que os itens se distribuam a contento entre mais de um fator. Entretanto, no caso de haver mais de um fator, as análises dos parâmetros dos itens a serem discutidos mais adiante devem ser feitas somente com os itens que pertencem a um dado fator. Assim, se AF mostrou que a sua matriz de in-

tercorrelações comporta dois fatores, você terá que fazer as análises dos itens divididos em dois grupos, como se fossem dois testes diferentes.

2 – Unidimensionalidade baseada na Teoria de Resposta ao Item

A TRI possui índices que analisam a adequação dos seus modelos de análise de itens. Se estes índices indicarem que o modelo utilizado for adequado, então fica demonstrado que o item é unidimensional. Há, novamente, uma legião de tais índices. Um desses índices consiste na análise dos resíduos, da qual falamos ao tratarmos da TRI (cap. 4). O problema contido neste modo de pensar consiste em que a análise da adequação do modelo deve ser feita depois que foi verificada a existência de unidimensionalidade da séric de itens analisados (Hambleton, Swaminathan, Cook & Gifford, 1978). De fato, analisando o TNVRA pela TRI, descobriu-se que o item 24, que a análise fatorial mostrou ser um item não unidimensional com os demais itens do teste, mostrou-se apesar disso perfeitamente adequado dentro dessa análise. Assim, a observação dos autores citados é válida: a TRI não demonstra a unidimensionalidade, ela a supõe.

3 – Unidimensionalidade baseada na análise fatorial Full Information

Bock e Aitkin (1981; Bock, Gibbons & Muraki, 1988) introduziram um novo método de análise fatorial de itens, baseado na TRI, o qual não requer o cálculo das intercorrelações entre os itens. Deram-lhe o nome de *Full Information Factor Analysis* – FIFA, porque trabalha com informações completas em lugar dos métodos de informação limitada ou sumarizada, tais como as correlações. Embora laborioso de um ponto de vista computacional, particularmente quando o número de itens for grande, este método evita a série de problemas que elencamos acima contra a análise fatorial tradicional e, no presente, parece ser o melhor método para decidir a unidimensionalidade de uma série de itens, tanto dicotômicos quanto politômicos (Bartholomew, 1980).

A FIFA trabalha com os padrões distintos de resposta ao item em vez de com as intercorrelações entre os itens, utilizando o modelo multifatorial de Thurstone (1947) baseado em estimativas de máxima verossimilhança marginal (*marginal maximum likelihood*) e no algoritmo EM (*expectation – maximization*) de Dempster, Laird e Rubin (1977). Vejamos o que é isso.

Os padrões de resposta referidos resultam, pela teoria, de vetores de resposta ao item. Um vetor de resposta ao item é expresso da seguinte forma, assumindo *r* fatores:

$$y_i = a_1\theta_1 + a_2\theta_2 + ... + a_r\theta_r + e_i \qquad (5.1)$$

onde,

y_i = processo de resposta dada ao item *i*

a_r = peso de cada fator no processo de resposta ao item

θ_r = os fatores envolvidos no processo de resposta (isto é, os traços latentes)

e_i = erros de medida

A FIFA analisa uma matriz quadrangular de respostas aos itens definidas pelos vetores de resposta acima descritos. O que é que ela analisa nesta matriz? Ela analisa os chamados padrões diferentes de resposta. O que é isso? Vejamos.

Se você aplica um teste de 30 itens, como o caso do TNVRA, a uma amostra de 1.000 sujeitos, você terá como resultado uma matriz quadrangular de 30 x 1.000. Nesta matriz cada sujeito tem uma sequência de 30 respostas; no caso de um teste de aptidão essa sequência é composta de 1 (acerto) e 0 (erro). Esta sequência é chamada de padrão de resposta do sujeito. Isso aparece como na ilustração abaixo:

Sujeito	Padrão de Resposta
1	1 0 1 0 0 1 0 1 1 1 0 0 1 1 0 1 0 0 1 0 1 1 1 0 0 1 0 0 1 0
2	1 1 0 0 1 1 0 1 0 0 1 0 1 0 0 1 0 1 1 1 0 0 1 1 0 1 0 0 0 0
3	1 0 1 0 0 1 0 1 1 1 0 0 1 1 0 1 0 0 1 0 1 1 1 1 0 0 1 1 0 1
4	1 0 1 0 0 1 0 1 1 1 0 0 1 1 0 1 0 0 1 0 1 1 1 0 0 1 0 0 1 0
...	...
1.000	1 0 1 0 0 1 0 1 1 1 0 0 1 1 0 1 0 0 0 0 1 0 1 0 0 1 0 1 1 0

A FIFA analisa todos os padrões que são diferentes. Assim, no caso dos 4 primeiros sujeitos da ilustração, há 3 padrões distintos de resposta, porque o padrão do sujeito 4 é igual ao do sujeito 1. A FIFA analisa estes padrões com a esperança de que eles todos possam ser reduzidos a um vetor de resposta aos itens no qual exista apenas um θ, ou seja, um único fator; o que iria demonstrar que todos os itens do teste são unidimensionais.

O número de padrões diferentes de resposta aumenta exponencialmente com o aumento do número de sujeitos, sobretudo quando o número de itens é grande, porque é difícil ocorrer que os sujeitos respondam do mesmo jeito todos os itens. Esse aumento de padrões distintos de resposta é o que torna as análises da FIFA extremamente laboriosas e, no limite, as torna quase impossíveis. De qualquer forma, existem softs para trabalhar a FIFA.

Mas a FIFA oferece grandes vantagens sobre os demais métodos de demonstração da unidimensionalidade, a saber:

1) ela leva em conta toda a informação empírica da aplicação do teste, não somente informações sintetizadas, como é o caso na AF que utiliza as correlações entre as variáveis;

2) ela consegue trabalhar o acerto dado ao acaso;

3) consegue tratar os dados omissos;

4) consegue contornar os problemas da matriz não positivo-definida, bem como os casos Heywood, que ocorrem frequentemente na análise fatorial.

Se você for trabalhar com a FIFA, vai ter que utilizar o pacote estatístico TESTFACT (Wilson, Wood & Gibbons, 1991).

B.2 – Dificuldade dos itens

Definição e cálculo da dificuldade

Na TCT, a dificuldade do item é definida em termos da percentagem (proporção) de sujeitos que dão respostas corretas ao item. Assim, um item que é respondido corretamente por 70% dos sujeitos é afirmado ser mais fácil que um que recebeu 30% de respostas corretas. Então, a dificuldade do item é dada pelo número (proporção) de acertos; cf. a tabela 5-3, onde esta informação para o TNVRA está dada na coluna encabeçada por

PC (proporção correta, isto é, número de acertos. Note que a tabela contém inúmeras informações que serão explicadas em seguida).

Tabela 5-3. Parâmetros dos itens do TNVRA

	TRI				TCT				Análise fatorial	DIF
Item	a	b	c	Res.	PC	r_{iT}	$r_{i\theta}$	r_{iC}	Carga	Viés
1	0,64	-0,62	0,13	1,45	68	0,37	0,37		0,44	
2	1,30	0,25	0,12	0,36	50	0,62	0,61		0,75	
3	1,44	0,29	0,11	0,72	48	0,66	0,64		0,79	
4	1,34	0,18	0,13	0,52	54	0,61	0,58		0,71	
5	0,60	0,59	0,12	1,70	46	0,37	0,33		0,44	
6	1,35	-0,48	0,12	0,31	69	0,63	0,64		0,79	
7	1,75	0,62	0,10	0,79	38	0,64	0,64		0,80	
8	1,24	-0,33	0,13	0,52	66	0,60	0,61		0,73	
9	1,74	0,19	0,10	0,77	50	0,71	0,70		0,85	
10	1,57	-0,12	0,13	0,54	61	0,67	0,66		0,80	
11	1,25	-0,47	0,14	0,97	70	0,59	0,58		0,72	
12	1,48	-0,08	0,11	0,71	58	0,68	0,68		0,82	
13	1,35	-0,46	0,12	0,48	69	0,62	0,63		0,77	
14	1,68	-0,06	0,10	0,77	57	0,71	0,71		0,85	
15	1,58	-0,88	0,12	0,45	79	0,58	0,64		0,78	
16	1,24	0,62	0,12	0,59	41	0,55	0,53		0,67	
17	1,52	0,55	0,12	0,78	43	0,59	0,57		0,71	
18	1,53	-0,04	0,10	1,58	56	0,69	0,70		0,82	
19	2,05	-0,44	0,12	0,53	69	0,70	0,72		0,87	
20	1,16	0,24	0,12	0,43	51	0,58	0,57		0,69	
21	1,56	0,28	0,10	1,08	48	0,68	0,68		0,83	
22	1,50	0,49	0,11	0,46	43	0,62	0,60		0,74	
23	1,93	-0,17	0,10	0,87	60	0,74	0,74		0,89	
24	1,36	1,46	0,16	1,26	28	0,27	0,25		0,29	
25	1,42	1,94	0,09	1,14	13	0,29	0,29		0,43	
26	1,12	0,60	0,12	0,44	42	0,54	0,52		0,65	
27	1,16	0,50	0,12	0,72	44	0,56	0,55		0,66	
28	0,94	-0,44	0,13	0,76	67	0,52	0,52		0,64	
29	1,11	1,66	0,13	0,60	24	0,30	0,27		0,36	
30	1,29	1,27	0,10	0,95	25	0,44	0,43		0,57	

A = discriminação; b = dificuldade; c = chute; PC = percentagem correta; r_{iT} = correlação item-total; $r_{i\theta}$ = correlação item-teta; r_{iC} = correlação item-critério

Na Psicometria Clássica, o conceito de dificuldade do item faz sentido somente no contexto de testes de aptidão, onde há respostas certas e erradas; não havendo, portanto, lugar para tal no contexto de testes cuja resposta depende da preferência do sujeito (testes de personalidade, por exemplo) e não de sua capacidade. De qualquer forma, a Psicometria tradicional tem tratado da dificuldade do item em testes de aptidão definindo-a como a percentagem de acertos (isto é, de fato se trata de índice de facilidade, pois quanto mais sujeitos acertam o item, mais fácil ele é), cuja fórmula é

$$ID = \frac{A}{N} \qquad (5.2)$$

Coeficiente ou índice de dificuldade do item

onde,

A: número de sujeitos que acertaram o item

N: número total de sujeitos que responderam o item.

A TRI, por outro lado, define dificuldade do item em termos do traço latente, do teta (θ), dizendo que esta dificuldade é diretamente proporcional ao nível ou tamanho de teta necessário para que um dado item possa ser acertado (testes de aptidão) ou aceito (testes de personalidade). Assim, um item é tanto mais difícil quanto maior for o tamanho do teta que o sujeito deve possuir para poder acertar ou aceitar o mesmo item. A TRI nomeia este parâmetro do item de *b* (cf. tabela 5-3) ou de limiar (*threshold, location*), porque ele é definido pela perpendicular, sobre a abscissa, da curva característica do item (CCI) no momento da inflexão, isto é, no ponto onde ocorre a probabilidade de 50% de acertar e 50% de errar o item (ou aceitar), como ficou esclarecido no capítulo 4. Desta forma, você vê que o conceito de dificuldade do item se aplica a qualquer tipo de testes, sejam de aptidão ou de preferência, pois o critério de dificuldade não é acertar ou errar o item e sim a magnitude do traço latente (qualquer traço latente) necessária para acertar ou aceitar o item. Veja, por exemplo, os seguintes três itens de afiliação (de um teste de personalidade):

a) Eu não detesto meus pais

b) Eu gosto de meus pais

c) Eu adoro meus pais

É óbvio que você não precisa de um grande teta de afiliação para aceitar o primeiro dos itens acima, enquanto para aceitar o terceiro você terá que possuir um teta muito grande de afiliação. Consequentemente, o primeiro item é fácil e o terceiro é difícil, porque a aceitação deste exige uma magnitude maior de teta de afiliação do que aquele. Expressar tal ocorrência como dificuldade pode parecer estranho a você, mas é a expressão utilizada na TRI e corresponde corretamente ao que este modelo entende por dificuldade do item.

A TRI expressa a dificuldade do item numa escala padrão, isto é, em escores z sob o nome de b (alguns autores chamam de p). A escala dos z varia de $-\infty$ a $+\infty$, mas na prática ela vai de -3 a +3, porque entre estes dois últimos extremos se situam mais de 99,73% dos casos[3]. Cf. a figura 5-6 para verificar como se descreve a dificuldade dos itens dentro da TRI.

a)

Item	b
1	-0,05
2	0,50
3	1,70

b)

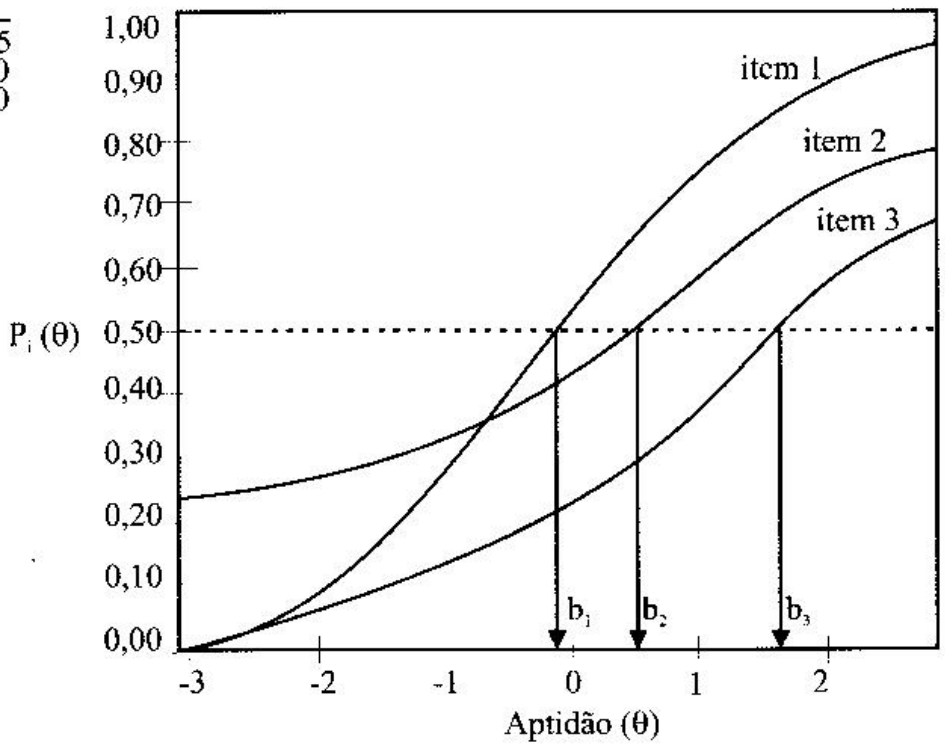


Figura 5-6. Expressão algébrica (a) e geométrica (b) da dificuldade do item

3. Esta escala dos z para fins práticos é bastante deselegante. No capítulo sobre a Normatização dos Testes serão indicadas maneiras de transformar esta escala em uma mais compreensível e elegante.

Relacionamentos da dificuldade com outros parâmetros psicométricos

No caso dos testes de aptidão, este índice ou coeficiente de dificuldade dos itens está diretamente relacionado com a média do teste, isto é, a média é igual à soma dos coeficientes de dificuldade dos itens do teste. A fórmula é

$$\overline{X} = \sum ID \qquad (5.3)$$

O exemplo da tabela 5-4 ilustra esta relação, em cujo corpo da tabela o 1 significa que o item foi acertado e 0 que foi errado.

Tabela 5-4. Cálculo do coeficiente de dificuldade dos itens e da média

	Itens						Escore Total
Sujeitos	1	2	3	4	5	6	T
1	1	1	0	1	0	0	3
2	1	0	1	1	1	0	4
3	0	1	1	0	0	0	2
4	1	1	1	1	1	1	6
5	1	0	0	0	0	0	1
ID	4/5	3/5	3/5	3/5	2/5	1/5	16

$$\overline{X} = \frac{\sum T}{N} = \frac{3+4+2+6+1}{5} = 3{,}2$$

$$\overline{X} = \sum ID = \frac{4}{5} + \frac{3}{5} + \frac{3}{5} + \frac{3}{5} + \frac{2}{5} + \frac{1}{5} = \frac{16}{5} = 3{,}2$$

Há igualmente uma relação direta entre o ID de um item e a sua variância, pois esta é $s_j^2 = pq$, onde p é a proporção de sujeitos que acertam o item e q dos que erram ($q = 1 - p$). Assim, a variância dos dados da tabela 5-3 será a soma dos produtos da multiplicação dos ID (p) e seus complementares (q = 1-p) dos 6 itens, a saber

$$s^2 = \left(\frac{4}{5} x \frac{1}{5}\right) + \left(\frac{3}{5} x \frac{2}{5}\right) + \left(\frac{3}{5} x \frac{2}{5}\right) + \left(\frac{3}{5} x \frac{2}{5}\right) + \left(\frac{2}{5} x \frac{3}{5}\right) + \left(\frac{1}{5} x \frac{4}{5}\right) = 1{,}28.$$

Correção para o chute

A dificuldade de um item pode ser afetada por respostas dos sujeitos que não são dadas em função de sua aptidão, mas simplesmente por "chute". Isto é, o sujeito responde, não em função de sua aptidão ou traço latente em questão, mas em função de outros fatores aleatórios ou outros traços latentes. Tal situação pode ocorrer, por exemplo em testes de aptidão, tal que o sujeito acerte o item por acaso, resultando em que o item se torna, erroneamente, menos difícil. No caso de testes de aptidão, onde este problema é mais óbvio, o chute ocorre sobretudo com os itens que são de fato mais difíceis, para os quais sujeitos de menor aptidão não conhecem a resposta correta e, então, a chutam e, às vezes, acertam. E este acerto é que é o problema. Assim, é preciso utilizar técnicas para controlar este tipo de resposta.

Na TCT, no caso de itens de múltipla escolha, o índice de dificuldade (ID) pode ser corrigido para levar em conta essas respostas corretas dadas por azar (o chute correto), utilizando a fórmula seguinte:

$$ID = \frac{A - \frac{E}{K-1}}{N} \qquad (5.4)$$

onde,

A: número de sujeitos que acertam o item

E: número de sujeitos que erram o item

K: número de alternativas de resposta ao item

N: número total de sujeitos.

Esta fórmula assume que, para cada k-1 respostas incorretas, há uma resposta correta conseguida por acaso. Essa suposição parece não satisfatória. Por exemplo, Lord (1968) dizia que sujeitos que possuem uma informação, pelo menos parcial, sobre um item ou têm uma informação errada sobre

o mesmo, não respondem a ele de forma aleatória, de sorte que nem todas as alternativas têm o mesmo apelo para o sujeito respondente.

A TRI estima o chute através do parâmetro *c*. Este parâmetro é definido pela assíntota da CCI: se ela cortar a ordenada acima do ponto 0, então houve chute, isto é, há respostas corretas por parte de sujeitos que não poderiam conhecer a resposta correta, já que seu nível de aptidão é baixo demais. Cf. por exemplo o item 2 da figura 5-1: sua CCI corta a ordenada em 0,20 para sujeitos que possuem um teta inferior a -3; então houve 20% de respostas corretas dadas por acaso a este item. Por que assim? Porque você vê que os 20% dos sujeitos que acertaram este item não têm aptidão (magnitude do teta) suficiente para poderem conhecer a resposta correta; de fato, o teta deles é extremamente baixo (menos do que -3), enquanto a dificuldade do item 2 corresponde a um teta de 0; assim, se, apesar disso, esses sujeitos acertam o item, só pode ser por acaso, isto é, chutaram a resposta e tiveram sorte (acertaram).

Outra forma de avaliar o chute pela TRI consiste no seguinte: ordene todos os itens por seu nível de dificuldade; se itens difíceis são respondidos por sujeitos de pouca habilidade, houve chute.

Também estas técnicas da TRI de avaliar o chute têm seus problemas. Wainer (1983) observa que a gente nunca sabe se o sujeito acertou por acaso ou por outras razões pessoais dele ou, mesmo, que o item esteja representando mais de um traço latente (é multidimensional) para certo tipo de populações e, assim, este sujeito pode ter acertado o item devido a um desses outros traços latentes que o item cobre e no qual o sujeito é forte, mesmo sendo fraco no traço latente dominante que o teste pretende medir; o autor dá um exemplo: num teste de conhecimento de drogas, um estudante indígena disse conhecer muito pouco sobre elas, mas acertou itens difíceis, que, examinados com maior atenção, percebeu-se que lidavam com mescalina, uma droga comum em cerimônias indígenas. Por causa destes problemas e outros, muitos pesquisadores não gostam de trabalhar com o modelo logístico de 3-parâmetros, onde se calcula também o chute. Infelizmente, este é ainda um problema não resolvido satisfatoriamente na TRI.

Avaliação crítica

O modo pelo qual a TCT calcula o ID contém um problema, pois ele depende diretamente da amostra de sujeitos sobre a qual ele é calcula-

do. Na verdade, o ID de um item pode mostrar ser o item muito difícil se os sujeitos da amostra forem de pouca aptidão e pode mostrar ser fácil se os sujeitos forem de habilidade superior (cf. figura 5-7). Assim, pela Figura 5-7, vê-se que o item *j* será fácil para a amostra B, que tem uma média superior, e difícil para a amostra A, cuja média é inferior à do B.

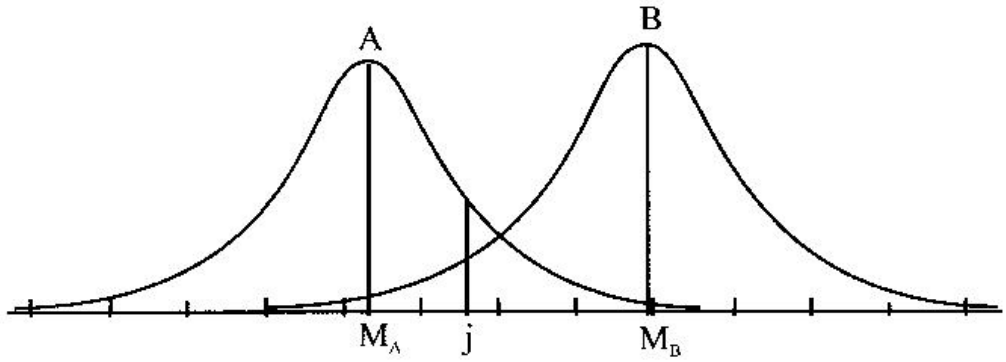


Figura 5-7. Distribuição da dificuldade dos itens em termos de duas amostras de aptidão diferente

Este problema, contudo, desaparece se a amostra de sujeitos for uma amostra representativa da população, pois neste caso a amostra possui os parâmetros da população e, portanto, qualquer amostra aleatória desta população irá produzir os mesmos ID para os itens.

Nível ideal de dificuldade dos itens de um teste

Pode-se, finalmente, perguntar ainda se existe um nível ideal de dificuldade para os itens de uma escala ou teste. A resposta a esta indagação depende da finalidade do teste. Se for desejado um teste para selecionar os melhores ou para determinar se um patamar 'x' de conhecimento foi atingido (como nos testes 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 for desejado 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 discriminariam 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, 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 para 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 e itens que ninguém acerta 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%, pois neste caso 50% dos sujeitos acertam e 50% erram, resultando 50 x 50 = 2.500 comparações possíveis, ao passo que um item com dificuldade de 30% teria 70% de erros e 30% de acertos, resultando num nível de 30 x 70 = 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 os itens o deverão, pois 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, ela pode ser dividida em cinco níveis de magnitudes: 0 a 20, 20 a 40, 40 a 60, 60 a 80 e 80 a 100, 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 (cf. figura 5-8, onde estão representados também os valores z que correspondem às divisórias de cada faixa).

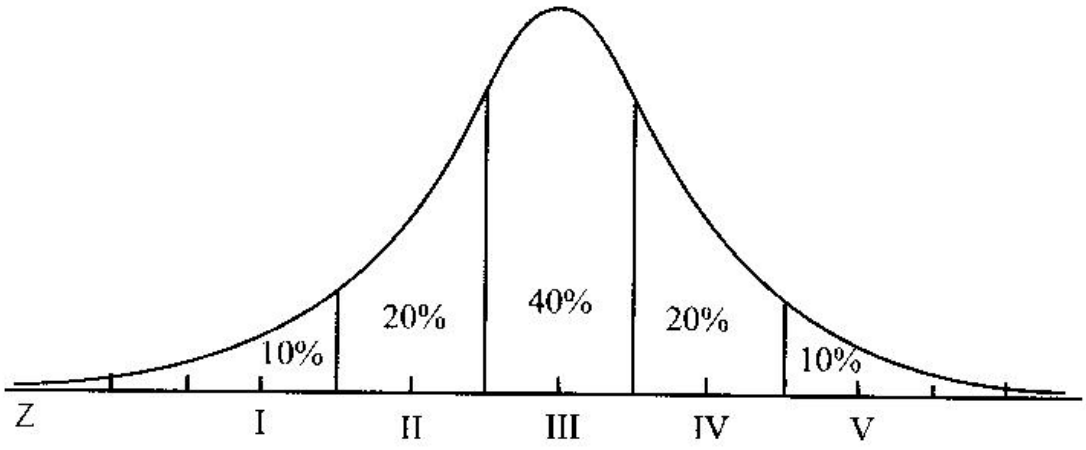


Figura 5-8. Distribuição percentual dos itens em 5 faixas de dificuldade

As observações feitas acima valem para testes construídos tanto pela TCT quanto pela TRI. No caso da TRI, entretanto, vem se popularizando mais e mais os chamados testes sob medida (*adaptive testing* ou *tailored testing*), em cujo caso as regras são diferentes, como falaremos sobre o assunto no capítulo 10.

Formas de apresentar a dificuldade dos itens

A TRI apresenta a dificuldade dos itens com o parâmetro *b*. A TCT, entretanto, desenvolveu várias maneiras para a apresentação da dificuldade dos itens. A forma discutida até aqui foi em termos da proporção de acertos (as percentagens). Assim, se os itens 1, 2 e 3 foram acertados por 30%, 40% e 50% dos sujeitos, sabemos que o item 1 é o mais difícil e o item 3 o mais fácil. Desta forma, a escala de dificuldade em termos de percentagens de acertos, que vai de 0 a 100, ou, em termos de proporções de acertos, que vai de p = 0,00 a p = 1,00, constitui uma escala ordinal. Isto porque a distância entre um índice de dificuldade de 10 e de 20 não é a mesma que há, por exemplo, entre os índices 50 e 60, como mostra a figura 5-9. Tal seria o caso somente se a distribuição dos índices fosse retangular sobre toda a amplitude da escala; na verdade, a distribuição é de percentagens que se distribuem sob curvas de tipo normal ou assimétrico (cf. capítulo 8 sobre normas percentílicas).

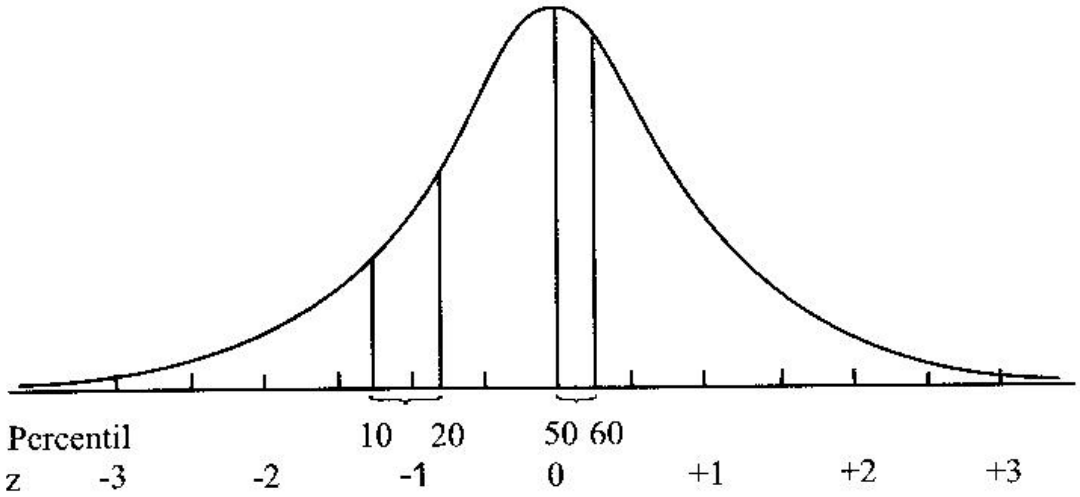


Figura 5-9. Distâncias entre diferentes pontos percentílicos

Entretanto, quando se puder assumir que a distribuição do traço medido pelo teste se distribui normalmente na população, então se pode construir uma escala intervalar para os índices de dificuldade dos itens, expressando estes índices em termos de escores *z*. A Teoria de Resposta ao Item (IRT) faz isso, utilizando um modelo logístico de análise. Mas mesmo dentro da Psicometria Clássica podemos utilizar deste estratagema, como faz o *Educational Testing Service* (ETS) de New Jersey. A escala utilizada pelo ETS é uma escala intervalar que vai de 1 a 25, a chamada escala Delta (Δ). Esta escala é simplesmente uma transformação da escala dos *z* da curva normal pela fórmula

$$\Delta = 13 + 4z.$$

Estes números estranhos da fórmula, isto é, 13 e 4, poderiam ser quaisquer outros. Mas a razão deles é a seguinte: O item que todos ou quase todos acertam vai produzir um escore *z* de -3σ na curva normal, o que corresponde a dizer que 99,87% dos sujeitos acertam o item, isto é, o item é extremamente fácil; um outro item que todo o mundo ou quase erra vai dar um *z* de +3σ, isto é, vai dar 0,13% de acertos ou 99,87% de erros do item, um item extremamente difícil. Assim, a escala de dificuldade, em termos de *z*, estende-se, na prática, de -3 a +3, passando por todas as decimais (de -3,00 até +3,00), obtendo +3,00 o item mais difícil e -3,00 o mais fácil. Como ela parece deselegante, isto é, ela apresenta decimais e valores negativos, o ETS inventou estes dois números (i.é, 13 e 4) para eliminar ambas as deselegâncias. Assim, a escala vai de 1 [13 + 4(-3) = 1] a 25 [13 + 4(3) = 25], passando por 13 [13 + 4(0) = 13] como indicativo do índice médio de

dificuldade dos itens. A grande vantagem desta escala (da escala *z* em geral) é que ela é uma escala intervalar, significando que as distâncias entre os índices de dificuldade Delta são iguais. Isto quer dizer que o item com Delta = 15 em relação ao que tem Delta = 14 tem um Delta a mais de dificuldade, o mesmo que entre um item com Delta = 5 em relação ao item com Delta = 4. Trata-se de uma vantagem para estatísticos ou para os iniciados; o profissional psicólogo irá, na prática, achar esta escala esdrúxula! Note que a TRI trabalha diretamente com a escala de *z*.

As duas maneiras de expressar o índice de dificuldade estão expressas na figura 5-10.

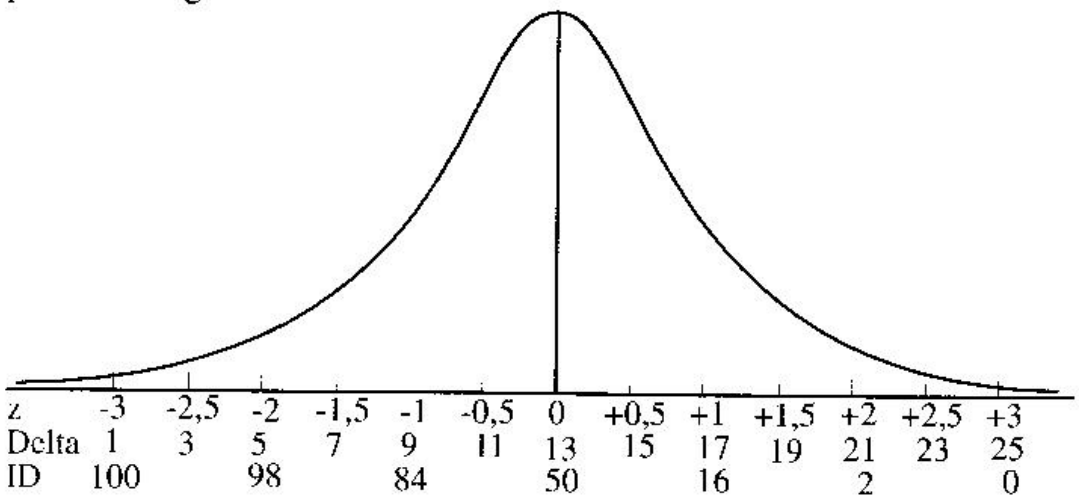


Figura 5-10. Distribuição da dificuldade dos itens em termos de ID, z e Delta

As escalas *z* e Delta estão expressas pelas distâncias na linha de base da curva normal, ao passo que a escala *p* ou percentagens ou ID está expressa na área sob a curva normal, onde a cada distância *z* corresponde uma percentagem de área diferenciada. Assim, um item respondido por 84% da amostra de sujeitos é um item fácil e tem um índice **ID** = 84, um **z** = -1 e um **Delta** = 9.

B.3 – Discriminação dos itens

A TCT define a discriminação do item como a sua capacidade de diferenciar sujeitos com escores altos no teste de sujeitos com escores baixos.

Desta conceituação de discriminação surgem duas formas de calcular estatisticamente o índice de discriminação: 1) a dos grupos-critério e 2) a correlação do item com o total dos itens.

3.2.1 – Grupos-critério

A dificuldade envolvida na tarefa de avaliar o poder discriminativo dos itens consiste na escolha dos sujeitos que servirão de base como grupos-critério que o item deve diferenciar. A escolha dos critérios para efetuar a análise da discriminação dos itens tem dependido, na prática, dos objetivos do teste. Assim existem critérios externos e critérios internos ao próprio teste cujos itens se quer analisar. *Critérios externos* para estabelecer os grupos-critério podem ser, por exemplo, sujeitos psiquiátricos e sujeitos não psiquiátricos para avaliar o poder de discriminação dos itens em testes psiquiátricos, ou sujeitos que tiveram êxito e sujeitos que fracassaram num curso de treinamento, ou, ainda, tipos de ocupações, etc. Enfim, trata-se de estabelecer grupos que se diferenciam em algum comportamento definido como relevante com referência aos objetivos do teste e verificar se os itens do teste são capazes de, individualmente, diferenciar estes grupos de sujeitos.

Utilizam-se também *critérios internos* ao próprio teste para definir estes grupos-critério. Tipicamente é escolhido o escore total no próprio teste para determinar os grupos extremos de sujeitos: grupo superior e grupo inferior. Em amostras grandes, selecionam-se os 27% superiores e os 27% inferiores para comporem os dois grupos (Kelley, 1939). Evidentemente, em amostras menores, este percentual deverá ser maior, visto que os grupos de comparação devem apresentar um número suficiente de sujeitos para permitir análises estatísticas válidas. De modo geral, algo em torno de 30% será adequado; contudo, em amostras normais e grandes é costumeiro se utilizar a "regra 27", como ficou sendo conhecida. Este procedimento necessita que se calcule primeiro o escore total de cada sujeito no teste. Para o grupo superior são, então, escolhidos os 27% de sujeitos com os maiores escores no teste e, para o grupo inferior, os 27% de sujeitos com os menores escores no teste.

Com base nestes grupos-critério pode-se calcular o índice de discriminação através da estatística D e do teste t de Student. Estas estatísticas analisam a diferença de percentagens ou de médias dos sujeitos que passaram (testes de aptidão, onde há respostas certas e erradas) ou aceitaram (testes de personalidade, atitude) o item no grupo superior vis-à-vis o grupo inferior, sendo estes dois os grupos-critério.

a) O índice D

É um dos mais fáceis para ser computado porque consiste simplesmente na diferença de percentagens de acertos no grupo superior e no grupo inferior, isto é, S – I ou, em inglês, U – L (ULI ou ULD). Cf. exemplo na tabela 5-5.

Tabela 5-5. Computação do índice D

Item	% dos que passaram		Índice D
	Grupo Superior	Grupo Inferior	
1	80	40	40
2	100	90	10
3	30	50	-20
4	55	55	0
5	75	40	35

O índice D tem que ser positivo e quanto maior for, mais discriminativo será o item. Obviamente, um D nulo ou negativo demonstra ser o item não discriminativo.

b) O teste t

Um índice de discriminação mais exato, embora mais laborioso de se conseguir, consiste na análise da diferença entre as médias obtidas pelo grupo superior e inferior. Neste caso, é necessário o cálculo das respectivas médias e de suas variâncias. O nível de significância do teste *t* pode ser verificado com exatidão em tabelas estatísticas próprias. Esta estatística faz uso das médias e das variâncias, portanto ela faz as exigências de que: (1) os escores constituam uma variável contínua e (2) ambos os grupos possuam variância, isto é, nenhum dos grupos pode apresentar variância 0.

A fórmula para o cálculo do teste *t* é a seguinte:

$$t = \frac{\overline{X}_s - \overline{X}_i}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_i^2}{n_i}}}$$

com graus de liberdade, $gl = n_s + n_i - 2$

onde,

$\overline{X}_s$ e $\overline{X}_i$ são as médias do grupo superior e inferior

s_s^2 e s_i^2 são as variâncias do grupo superior e inferior

n_s e n_i é o número de sujeitos nos dois grupos.

Exemplo: 100 sujeitos responderam a um teste de 5 itens numa escala de 7 pontos. Os itens tiveram o índice de discriminação (teste *t*) segundo tabela 5-6:

Tabela 5-6. Discriminação dos itens pelo cálculo do teste *t*

Item	Média do Grupo		Variância do Grupo			
	Superior	Inferior	Superior	Inferior	t	p
1	5,5	3,2	1,50	0,99	7,56	< 0,01
2	6,2	5,0	2,00	1,50	3,33	< 0,01
3	5,3	3,0	1,45	1,00	7,64	< 0,01
4	3,0	5,0	1,05	2,50	-5,52	< 0,01*
5	3,3	3,0	1,00	0,95	1,12	ns
n	27	27				

Os quatro primeiros itens são discriminativos, mas o #4 o é ao avesso, isto é, o grupo inferior teve média maior que o grupo superior, demonstrando ser este item obviamente inadequado. O item 5 não é discriminativo, pois não consegue diferenciar claramente os dois grupos-critério.

3.2.2 – Correlação item total

Existem dezenas de técnicas estatísticas de correlação para estabelecer o índice de discriminação do item (Anastasi, 1988), as quais produzem basicamente resultados similares (Oosterhof, 1976). Algumas apenas serão aqui abordadas por serem as mais populares.

a) Correlação ponto-bisserial

Esta é a correlação de Pearson utilizada quando uma das variáveis (item) é dicotômica, coisa que ocorre com testes de aptidão onde a resposta é certo ou errado. Esta é a correlação apresentada para o TNVRA na tabela 5-3. A fórmula é

$$r_{pb} = \frac{\overline{X}_A - \overline{X}_T}{s_T}\sqrt{\frac{p}{q}} \qquad (5.5)$$

onde,

$\overline{X}_A$: média no teste dos sujeitos que acertam o item

$\overline{X}_T$: média total do teste

s_T : desvio padrão do teste

p: proporção de sujeitos que acertam o item

q = 1 – p.

Exemplo: 5 sujeitos responderam a um teste de 6 itens, sendo as respostas conforme a tabela 5-7, onde se calcula também a correlação ponto-bisserial para o item 3.

Tabela 5-7. Cálculo do índice de discriminação do item 3 via correlação ponto-bisserial

	Itens						Escore Total	
Sujeitos	1	2	3	4	5	6	T	T – j
1	1	1	0	1	0	0	3	3
2	1	0	1	1	1	0	4	3
3	0	1	1	0	0	0	2	1
4	1	1	1	1	1	1	6	5
5	1	0	0	0	0	0	1	1

1 = acertou o item, 0 = errou o item; j, no caso, é o escore no item 3.

Para o cálculo do índice de correlação ponto-bisserial do item 3, o escore total a ser utilizado é aquele do qual foi retirada a resposta dada ao próprio item 3, do contrário ele próprio contribuiria espuriamente para sua própria correlação com o restante dos itens. Esta precaução não é tão grave quando o número de itens no teste é grande ($n \geq 30$). Assim, o escore total a ser usado é o T-j. Calculando, dá

$\overline{X}_A = \frac{3+1+5}{3} = 3,0$ (média dos que acertaram o item 3, que são os sujeitos 2, 3 e 4)

$\overline{X}_T = \frac{3+3+1+5+1}{5} = 2,6$ (média total do teste)

$s_T^2 = \frac{3^2+3^2+1^2+5^2+1^2}{5} - 2,6^2 = 2,24$

$s_T = \sqrt{2,24} = 1,50$ (desvio padrão do teste)

$p = \frac{3}{5} = 0,60$ (proporção de sujeitos que acertaram o item 3)

$q = 1 - 0,60 = 0,40$ (proporção de sujeitos que erraram o item 3)

$r_{pb} = \frac{3,0 - 2,6}{1,50} \sqrt{\frac{0,60}{0,40}} = 0,33$.

A correlação é suficientemente elevada, indicando que o item se apresenta como razoavelmente discriminativo.

b) Correlação bisserial

Esta correlação é uma estimativa da correlação de Pearson e é utilizada na situação na qual as variáveis correlacionadas são contínuas, mas uma delas (no caso, o item) foi artificialmente reduzida a duas categorias (dicotomizada). A fórmula é

$$r_b = \frac{\overline{X}_A - \overline{X}_T}{s_T} \times \frac{pq}{y} \qquad (5.6)$$

onde,

– os símbolos são os mesmos da fórmula anterior

– y é a ordenada na curva normal no ponto de divisão dos segmentos que contêm as proporções p e q dos casos (cf. figura 5-11).

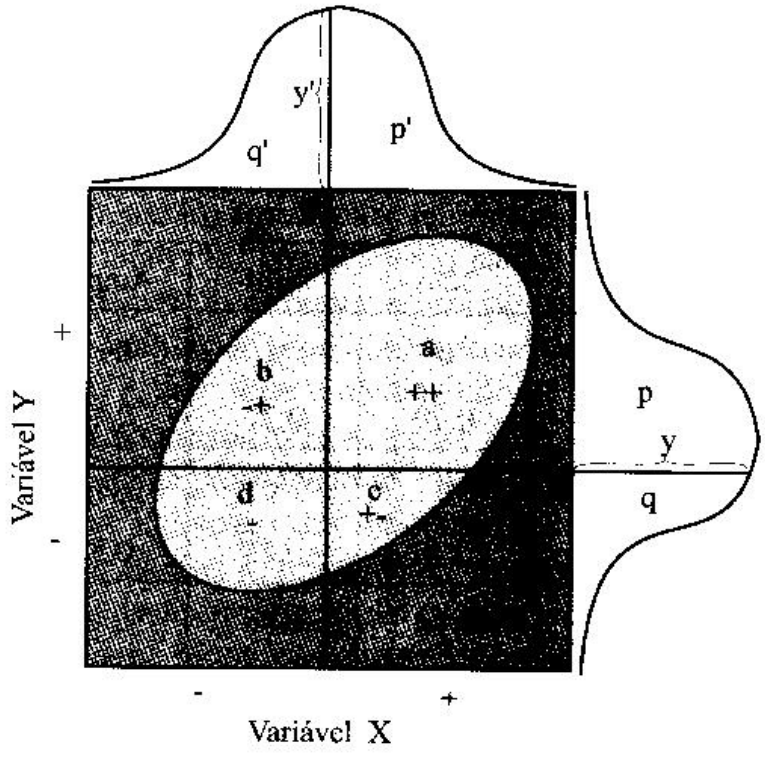


Figura 5-11. Distribuição bivariada normal teórica de casos em 4 categorias de 2 variáveis dicotomizadas.

Esta fórmula pode dar valores maiores do que 1 quando a distribuição não for normal (distribuição platicúrtica ou bimodal). Neste caso, seria melhor utilizar a correlação ponto-bisserial que tem a seguinte relação com a bisserial:

$$r_{pb} = r_b \frac{y}{\sqrt{pq}} \qquad \text{e} \qquad r_b = r_{pb} \frac{\sqrt{pq}}{y} \qquad (5.7)$$

c) Correlação phi (Φ)

Este coeficiente de correlação é utilizado quando as duas variáveis a correlacionar são genuinamente dicotômicas. A fórmula é

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} \tag{5.8}$$

Exemplo: A tabela 5-7 mostra como calcular o coeficiente phi.

Tabela 5-8. Cálculo do coeficiente phi

		Item 1		
		Não	Sim	Ambos
Item 2	Sim	**b** 0,20	**a** 0,27	**b+a** 0,47
	Não	**d** 0,30	**c** 0,23	**d+c** 0,53
	Ambos	**b+d** 0,50	**a+c** 0,50	 1,00

Sim e Não podem ser Acertou e Errou o item

$$\phi = \frac{0,27 \text{ x } 0,30 - 0,20 \text{ x } 0,23}{\sqrt{(0,27+0,20)(0,27+0,23)(0,20+0,30)(0,23+0,30)}} = 0,14$$

d) Correlação tetracórica

A correlação tetracórica é utilizada quando duas variáveis contínuas e normalmente distribuídas foram artificialmente reduzidas a duas categorias (dicotomizadas). A fórmula é

$$r_t = \frac{ad - bc}{yy' N^2} \tag{5.9}$$

onde,

– *a* até *d* são as caselas como na Tabela 5-8

– *N* é o número de sujeitos

– *y* e *y'* são as ordenadas como na Figura 5-11.

3.2.3 – Avaliação crítica do cálculo da discriminação pela TCT

O cálculo do índice de discriminação com base no escore total do teste, seja para permitir estabelecer os grupos-critério com o próprio teste ou para estabelecer a correlação entre o item e o escore no teste, apresenta um problema teórico. Na verdade, procura-se analisar a adequação do item (em termos de discriminação) baseada nas informações obtidas de todo o elenco de itens (escore total). Tal procedimento parece incongruente, dado que a adequação dos demais itens também está por ser demonstrada, inclusive a esta altura das análises do teste ainda não se sabe se os itens do teste são homogêneos, isto é, se o teste é unidimensional (medindo um único construto), suposição necessária para se poder obter um escore total. Tenta-se resolver este problema procedendo-se a uma análise fatorial dos itens antes da própria análise individual dos mesmos, para se certificar, pelo menos, que todos os itens estejam se referindo ao mesmo fator. Além disso, como a base dos cálculos da discriminação são os grupos-critério, acontece que para o caso de itens muito fáceis, os quais todos os sujeitos acertam, ou de itens muito difíceis, os quais todo o mundo erra, um dos grupos-critério desaparece, tornando o cálculo da discriminação inviável. Assim, para itens muito fáceis ou muito difíceis, a informação sobre a discriminação dos mesmos dada pela TCT é não confiável (normalmente o índice de discriminação, nesses casos, se aproxima de zero, provocando a eliminação errônea de tais itens por falta de discriminação).

A TRI tem outra maneira de definir e calcular a discriminação dos itens. A definição é a seguinte: Discriminação se refere ao poder de um item em diferenciar sujeitos com magnitudes diferentes de traço do qual o item constitui a representação comportamental. Quanto mais próximas forem as magnitudes do traço que o item puder diferenciar, mais discriminativo ele é. O cálculo deste índice é feito pelo algoritmo da função da curva característica do item e vem expresso com a letra *a* (veja na tabela 5-3) que indica o ângulo de incidência da CCI no momento da inflexão (*slope*), isto é, no momento em que ela corta o ponto de probabilidade de

50% (veja capítulo 4). Assim, poder-se-ia dizer que discriminação se refere ao poder que o item possui de diferenciar sujeitos com magnitudes próximas do traço a que se refere. Você vê que, no caso da TRI, a dificuldade do item não interfere no cálculo da sua discriminação, permitindo uma estimação adequada desse parâmetro também para itens muito fáceis ou muito difíceis.

B.4 – Validade dos itens

Fala-se de validade do item para designar o fato do mesmo estar relacionado com aquilo que pretende medir. Na TCT se diz que o item é válido se tiver alta correlação com o critério; na TRI, se ele for uma representação adequada do traço latente. Assim, você pode prever que há maneiras diferentes de calcular a validade do item. Na tabela 5-2, essas informações sobre o TNVRA estão dadas na coluna r_{iC} (correlação entre o item e o critério) para a TCT; para o caso da TRI, a coluna $r_{i\theta}$ (correlação entre o item e o teta) é uma aproximação desse parâmetro, apenas que o teta ali utilizado é uma transformação do escore total da TCT. Para a validade do item procurada pela TRI, é a carga da análise fatorial que melhor exprime essa informação (logo mais falaremos desta análise).

4.1 – Correlação item e critério

No caso da TCT, o cálculo da validade do item se faz através da correlação entre o item e o critério, sendo este um teste paralelo ao teste no qual o item, que está sendo avaliado, se encontra inserido; ou, melhor, a medida do desempenho numa situação para a qual o teste foi construído para predizer. Imagine o seguinte: você construiu um teste para predizer o desempenho de engenheiros mecânicos. Nesse caso, o critério é o desempenho dos mecânicos na sua situação de trabalho e a medida desse desempenho constitui o critério contra o qual os itens do seu teste serão avaliados. Por exemplo, os escores de 5 sujeitos no critério e suas respostas ao item *j* são os apresentados na tabela 5-9. Pergunta-se se o item tem um bom coeficiente de validade. Basta fazer a correlação entre os escores no critério com as respostas dadas ao item pelos 5 sujeitos. Vejamos.

Tabela 5-9. Cálculo do coeficiente de validade do item na TCT

Sujeito	Critério	Item j	c	j	c^2	j^2	cj
1	30	1	9	0,4	81	0,16	3,6
2	25	1	4	0,4	16	0,16	1,6
3	20	0	-1	-0,6	1	0,36	0,6
4	10	0	-11	-0,6	121	0,36	6,6
5	20	1	-1	0,4	1	0,16	-0,4
Soma	105	3			220	10,20	12,0
Média	21	0,6					

$$s_C^2 = \frac{220}{5} = 44{,}00 \text{ sendo } s_C = 6{,}63; \quad s_j^2 = \frac{1{,}20}{5} = 0{,}24, \text{ sendo } s_j = 0{,}49$$

$$r_{jC} = \frac{\Sigma cj}{N s_C s_j} = \frac{12}{5 \times 6{,}63 \times 0{,}49} = 0{,}74$$

c: desvio dos escores no critério; j: desvio da resposta no item

A correlação do item *j* com o critério é muito alta, indicando que o item possui boa validade.

4.2 – A carga fatorial

O tipo de validade calculada na TCT corresponde à validade preditiva, isto é, a capacidade do item predizer o critério. Entretanto, se você quiser saber se o item é um bom representante comportamental de algum traço latente, você irá utilizar ou a carga fatorial dada pela análise fatorial ou a curva de informação do item da TRI.

A carga fatorial de um item no fator mostra que percentual de parentesco ou de covariância existe entre o item e o fator. Esta covariância vai de 0% a 100%. Quanto maior for a covariância, maior é a validade do item, porque maior será sua representatividade do fator, sendo este o traço latente e o item sua representação empírica. Esta é a validade de construto do item. Para ver que percentual de covariância existe entre o item e o tra-

ço latente, basta elevar a carga fatorial ao quadrado e multiplicar por 100, que dá o chamado coeficiente de determinação. Por exemplo, a carga fatorial do item 7 do TNVRA é 0,80 (veja tabela 5-3); assim, o coeficiente de determinação será $0{,}80^2$ x 100 = 64, significando que há uma covariância de 64% entre o item e o teta, o que representa que o item é um excelente representante comportamental do referido teta.

4.3 – Função de informação do item

Além disso, a curva de informação do item acrescenta, a esta informação dada pela análise fatorial, o nível do teta para o qual este item traz a maior informação. Esta análise é feita através da função de informação do item da TRI. Este é um poderoso método para descrever itens, bem como para selecionar itens, pois permite analisar quanto um item traz de informação para a medida da aptidão.

A fórmula da função é a seguinte:

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \qquad (i = 1, 2, \ldots\ldots, n) \qquad (5.10)$$

onde,

$I_i(\theta)$ é a "informação" fornecida pelo item *i* ao nível da aptidão θ

$P_i'(\theta)$ é a derivada de $P_i(\theta)$ com relação a θ

$P_i(\theta)$ é a CCI e

$Q_i(\theta) = 1 - P_i(\theta)$.

No caso do modelo logístico de três parâmetros, a equação se simplifica para $I_i(\theta) = \dfrac{2{,}89a_i^2(1-c_i)}{[c_i + e^{1{,}7a_i(\theta-b_i)}][1+e^{-1{,}7a_i(\theta-b_i)}]^2}$

Esta equação mostra a importância que têm os três parâmetros sobre o montante de informação do item, como salienta a figura 5-12. Na verdade, a informação

1) é maior quando b_i se aproxima de θ

2) é maior quanto maior for o a_i e

3) aumenta com a diminuição de c_i para 0.

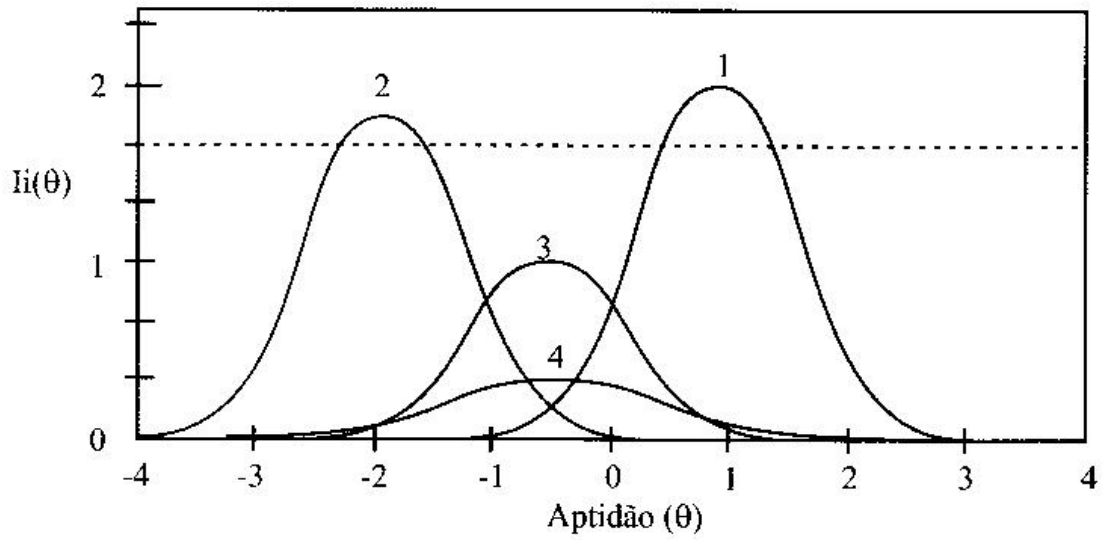


Figura 5-12. Função de informação para quatro itens

Observa-se na figura 5-12 que o item 1 fornece a maior informação a $\theta = 1{,}0$ e o item 2 a $\theta = -1{,}9$. O item 3 dá bem menos informação que os itens 1 e 2, porque o $c_3 > 0$, mas dá mais informação do que estes itens ao nível de $\theta = 0$. O item 4, devido ao grande c ($c_4 = 0{,}15$), tornou-se inútil no teste, pois não produz qualquer informação a mais do que a já produzida pelos outros itens 1, 2 e 3 a qualquer nível de θ.

B.5 – Algumas relações entre parâmetros dos itens e do teste

Parece importante salientar algumas das relações estatísticas que se encontram entre os parâmetros dos itens e os parâmetros de fidedignidade e validade do teste do qual eles fazem parte. Seguem algumas instâncias:

1) A *variância do teste* depende da capacidade discriminativa dos seus itens. Se estes não discriminam (isto é, todos os sujeitos obtêm o mesmo escore nos itens), então a variância do teste é 0. A fórmula é:

$$s_T = \sum s_j r_{jT} \qquad (5.11)$$

onde,

s_j : variância (desvio padrão) do item j

r_{jT} : índice de discriminação do item j

No caso dos itens dicotômicos, a variância do item é

$s_j^2 = p_j q_j = p_j(1-p_j)$.

Neste caso, a fórmula 5.10 será

$$s_T = \sum \sqrt{p_j(1-p_j)}\, r_{jT} \qquad (5.12)$$

Os termos $s_j r_{jT}$ e $\sqrt{p_j(1-p_j)}\, r_{jT}$ são conhecidos como sendo o índice de *fidedignidade do item j.*

2) A *fidedignidade do teste* pode ser expressa em termos da variância dos itens e de seus índices de discriminação pela fórmula seguinte:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum s_j^2}{\left(\sum s_j r_{jT}\right)^2}\right) \qquad (5.13)$$

ou, se o teste for composto de itens dicotômicos:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum p_j(1-p_j)}{\left(\sum r_{jT}\sqrt{p_j(1-p_j)}\right)^2}\right) \qquad (5.14)$$

3) *Validade do item* é expressa pela correlação entre o item e um critério, isto é, r_{jC}, utilizando uma das fórmulas apresentadas no tópico discriminação dos itens (ponto-bisserial, bisserial, phi, etc.). Que é este critério? No caso da TCT, a diferença entre o índice de fidedignidade do item (r_{jT}) e o índice de validade do mesmo (r_{jC}) é o fato de que a correlação é calculada com o próprio teste (T) do qual o item é parte no caso da

fidedignidade ou com outro teste (C) do qual o item não é parte no caso da validade. Isto faz sentido somente se C mede a mesma coisa que o T, isto é, trata-se de testes paralelos. No caso da TRI, o C é o teta (θ).

4) *Validade do teste* pode ser expressa em termos dos índices de discriminação e de validade de seus itens. A fórmula é:

$$r_{TC} = \frac{\sum s_j r_{jC}}{\sum s_j r_{jT}} \text{ ou, com itens dicotômicos } r_{TC} = \frac{\sum r_{jC}\sqrt{p_j(1-p_j)}}{\sum r_{jT}\sqrt{p_j(1-p_j)}} \qquad (5.15)$$

onde,

o somatório é feito sobre os itens (j = 1, 2, ..., n)

r_{jC} : índice de validade do item

r_{jT} : índice de discriminação do item

Assim, a validade do teste é o quociente da divisão da validade do item pela sua discriminação. De fato, a validade do teste depende de três parâmetros dos itens:

– dificuldade (p_j)

– discriminação (r_{jT})

– validade (r_{jC}).

Aqui aparece mais uma sinuca (chamam de paradoxo!) em que nos põem considerações puramente estatísticas em ciências. Na verdade

a) pela fórmula 5.12 a fidedignidade do teste (alfa, isto é, a consistência interna) aumenta com o aumento dos índices de discriminação dos itens

b) pela fórmula 5.13 a validade do teste diminui com o aumento dos índices de discriminação dos itens.

Que deve, então, o cientista fazer diante disto? Deve abandonar a estatística e se decidir pelo seu interesse ou objetivos da pesquisa: se ele quer teste mais válido ou mais fidedigno! Felizmente, estes não são os únicos índices de avaliar a validade dos itens e do teste, como veremos no capítulo sobre a validade.

B.6 – Vieses de resposta

Independentemente da qualidade dos itens ser boa, a resposta aos mesmos pode ser desvirtuada por fatores relativos ao sujeito que a eles reage. Estes vieses na resposta falseiam os dados, introduzindo correlações espúrias, mesmo em se tratando de bons instrumentos psicológicos. Podemos classificar estes erros em três categorias em termos de suas causas: cultura/nível socioeconômico, resposta aleatória, e resposta estereotipada.

a) A *cultura* como causa de erros de resposta se relaciona ao problema da transferência de instrumentos psicológicos para outras populações para as quais eles não foram especificamente construídos e validados. É o caso da utilização destes instrumentos para minorias e o caso da adaptação dos mesmos a outras culturas (tradução de testes).

O problema do uso dos testes com minorias tem atraído grande atenção nos Estados Unidos, sobretudo com a minoria negra, em seguida aos trabalhos de Jensen (1969, 1980). A Teoria da Resposta ao Item (TRI) também vem se preocupando com esta questão no contexto do uso de instrumentos para estudos transculturais (Hambleton, Swaminathan & Rogers, 1991). O problema que se observa ali é, sobretudo, a dificuldade relativa de certos itens para grupos de indivíduos com tradições culturais e de experiência diferentes das dos grupos para os quais os testes foram elaborados. Isto significa que pessoas de habilidades similares num dado construto psicológico, mas de culturas diferentes, apresentam diferentes probabilidades de êxito no teste.

Vários métodos estatísticos foram apresentados para lidar com este problema, chamado viés do item (*item bias*) e do teste (*test bias*), salientando-se o enfoque que analisa a proporção dos sujeitos que respondem corretamente o item em cada grupo – o método delta (Angoff & Ford, 1973; Angoff, 1982) e o DIF da psicometria moderna (Ironson, 1982; Hambleton, 1991; Ellis, 1991; Jackson, 1991).

A técnica de Angoff (1982) consiste em transformar as percentagens de acertos nas duas populações em valores delta[4] e plotá-los em coordenadas cartesianas. No caso das populações serem similares, esperam-se altas correlações entre as respostas dos sujeitos de ambas as amostras,

4. Sobre os escores delta, veja tópico sobre a dificuldade dos itens neste mesmo capítulo.

isto é, os itens se apresentam com dificuldades similares, resultando em um agrupamento dos itens em cima ou ao longo da linha de 45° que passa pelo ponto de origem das coordenadas. Quando estas populações, entretanto, forem culturalmente diferentes, o índice geral de dificuldade dos itens pode aparecer mais forte numa que na outra. Neste caso, os valores deltas não se agrupam junto à linha de 45°, mas os pontos (que definem os itens) aparecem mais afastados desta linha, isto é, aparece uma série de itens longe da linha, parecendo estranhos aos demais (*outsiders*). Estes itens são mais difíceis para a amostra de sujeitos onde eles aparecem plotados, como, por exemplo, o item assinalado por * na figura 5-13, o qual se apresenta bem mais difícil para a segunda amostra de sujeitos.

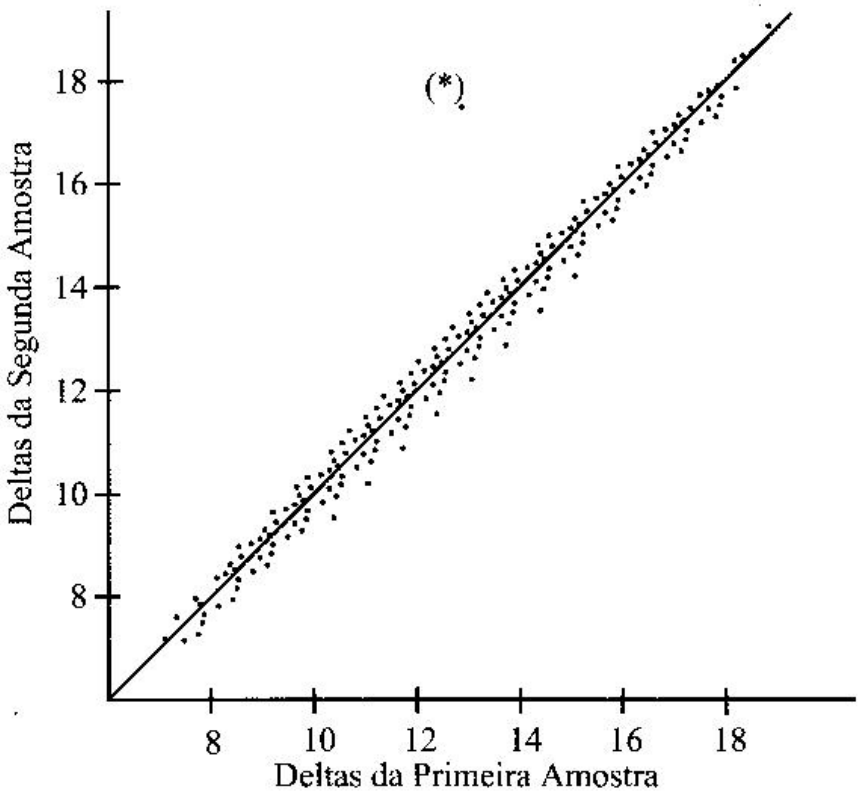


Figura 5-13. Distribuição hipotética da dificuldade dos itens em amostras de culturas diferentes

Além da representação gráfica sugerida na figura 5-13, Angoff e Ford (1973) propõem o cálculo de um índice geral das distâncias dos itens em relação ao eixo principal da elipse gerada pelos próprios itens quando aplicados a duas amostras distintas de sujeitos, como no caso da figura 5-8. Esta técnica implica no seguinte:

O eixo principal é dado por $Y = aX + b$

onde ,

$$a = \frac{s_y^2 - s_x^2 + \sqrt{(s_y^2 - s_x^2)^2 + 4r_{xy}^2 s_x^2 s_y^2}}{2r_{xy} s_x s_y}$$

$$b = \overline{Y} - a\overline{X}$$

O índice de distância é dado por:

$$d = \frac{aX_j - Y_j + b}{\sqrt{a^2 + 1}}$$

onde,

X_j : valor delta do item *j* no grupo X

Y_j : valor delta do item *j* no grupo Y

a e *b* são dados pelas fórmulas acima citadas.

Calculadas as distâncias *d* de todos os itens, pode-se calcular a média delas e descartar aqueles itens cujo valor de distância se afastam de um modo significativo desta média.

Note que o eixo principal não precisa cair nos 45° das coordenadas cartesianas, pois um teste pode de fato ser mais difícil para uma amostra que para outra. Contudo, o que mostra existir viés de item é este cair fora da elipse produzida pelos itens ao longo do eixo (ditos *outsiders*), como é o caso do item assinalado com * na figura 5-8.

A dificuldade com esta análise consiste no fato de que itens bem discriminativos tendem a se mostrar *outsiders* e, com isso, correrem o risco de serem eliminados como desviados. Outro problema consiste no fato de que a técnica supõe que os outros itens (isto é, a maioria) não sejam enviesados. Agora, se todos os itens forem enviesados, a técnica mostraria como enviesados somente aqueles cujos vieses destoassem do viés geral dos outros. Quer dizer que ela captaria itens cujo viés é simplesmente diferente do viés dos outros.

Estes problemas são evitados com o uso do enfoque da TRI para analisar os itens em termos de desviados quando aplicados a uma população culturalmente diferente da original, dado que ela analisa os parâmetros de cada item independentemente uns dos outros. Esta técnica permite analisar a equivalência dos itens quando aplicados a populações culturalmente distintas, identificando os itens que não apresentam tal equivalência, isto é, itens que apresentam um funcionamento diferencial (*differential item functioning* – DIF – Ellis, 1991). Para entender melhor esta história do DIF, imagine o seguinte: um teste foi aplicado a duas amostras diferentes de sujeitos e os escores médios foram diferentes. Essa diferença de médias pode ser devida ao fato de que realmente as duas amostras são diferentes na magnitude do traço que o teste mede. Essa diferença é chamada de *impacto*, porque ela é verdadeira e mostra uma diferença real que existe na habilidade das duas amostras. Entretanto, a diferença pode também ter sido provocada por artefatos do teste, tais como o tipo ou formato dos itens, que expressam modos de falar e representar as coisas diferentes nas duas amostras. Por exemplo, falar de avião e carro para um pigmeu das florestas africanas para se referir a voar e correr provavelmente não diz o que significa para um ocidental; neste caso, seria mais adequado falar de aves e outras criaturas ou objetos familiares ao pigmeu. Aplicar a um pigmeu testes do tipo mencionado faz com que eles tenham escores inferiores, simplesmente porque a representação comportamental do traço latente que o teste pretende medir não é adequada. Neste caso, em que um teste se apresenta em formas típicas de uma cultura, sendo estranhas para outra cultura, os resultados diferentes no teste são devidos ao DIF, porque não espelham diferenças reais de habilidade entre as culturas ou grupos e sim diferenças na representação do traço latente em comportamentos ou símbolos não familiares a um dos grupos. Evidentemente, neste caso, ao se utilizar um mesmo teste para grupos diferentes, deve-se corrigir o viés que prejudica um dos grupos. Como fazer isso?

O problema pode ser montado da seguinte maneira: existe mais de um grupo de sujeitos (grupo) que responde a um teste, o qual produz um escore em cima de itens respondidos num tipo de categoria (no caso de testes de aptidão, a categoria é certo ou errado ou 0 e 1). Assim, os dados a serem analisados se apresentam numa tabela de frequências de três entradas: Escore x grupo x categoria, como mostra a tabela 5-10.

Tabela 5-10. Tabela de frequência para análise do DIF

		Grupo		
Item	Categoria	Referência	Focal	Total
1	0	R10	F10	N10
	1	R11	F11	N11
2		frequências		
3				
...				
n	0	Rn0	Fn0	Nn0
	1	Rn1	Fn1	Nn1
Total	0	NRn0	NFn0	Nn0
	1	NRn1	NFn1	Nn1

Onde, grupo

- referência: grupo avaliado contra o grupo focal
- focal: grupo de base em função do qual é avaliado o grupo de referência
- NRn0: frequência de respostas erradas (0) no grupo de referência para o item *n*, etc.

Existe no mercado uma gama enorme de técnicas com respectivos softs para avaliar o DIF. Esta é, aliás, uma das áreas atualmente mais pesquisadas, onde as opiniões divergem muito. As técnicas giram, geralmente, em torno de vários métodos, quais sejam (Green, 1994),

- Métodos do qui quadrado (χ^2)
- Métodos de padronização
- Método de Mantel-Haenszel
- Métodos da TRI
- Variações entre estes métodos.

Quem quiser trabalhar com o DIF deverá consultar manuais especializados na área (cf. Bibliografia). Aqui vamos apenas exemplificar alguns desses métodos de análise do DIF (seguindo Green, 1994).

Métodos do χ^2

Para analisar o DIF de cada item, pode-se usar uma tabela de contingência do tipo da tabela 5-11.

Tabela 5-11. Tabela de frequência para análise do DIF

	Escore no item j		
Grupo	1 (acerto)	0 (erro)	Total (escore)
Referência	$\mathbf{A_j}$	$\mathbf{B_j}$	N_{rj}
Focal	$\mathbf{C_j}$	$\mathbf{D_j}$	N_{fj}
Total	N_{1j}	N_{0j}	$N_{.j}$

A fórmula para calcular um χ^2 do teste geral é a seguinte:

$$\chi^2 = \frac{N_{.j}\left(A_j D_j - B_j C_j\right)^2}{N_{1j} N_{0j} N_{rj} N_{fj}} \qquad (5.16)$$

Este qui-quadrado trabalha com T graus de liberdade (Shepard, Camilli & Williams, 1984).

Os procedimentos baseados no qui-quadrado têm sido muito criticados, porque (1) fazem combinações de dados (para evitar frequências baixas nas caselas) nem sempre justificadas e (2) demonstram pouco poder de decisão. Por isso, eles estão sendo substituídos pelos procedimentos de Mantel-Haenszel e pelos métodos de padronização.

Métodos de padronização

Estes métodos são uma transformação dos deltas de Angoff, ou seja, um delta padronizado, cuja fórmula é (Green, 1994; Dorans & Holland, 1992):

$$D_p = p_f - \hat{p}_f \qquad (5.17)$$

Onde,

$$p_f = \frac{\sum_j n_{fj} p_{fj}}{\sum_j n_{fj}}; \quad \hat{p}_f = \frac{\sum_j n_{fj} p_{rj}}{\sum_j n_{fj}}.$$

O procedimento de Mantel-Haenszel

Este método se aplica em casos de dois grupos e dados dicotômicos. Os dados são montados numa tabela de dupla entrada, encabeçadas as entradas por grupo (referência vs. focal) e por resposta ao item (certo vs. errado), conforme tabela 5-8. A hipótese a ser testada é de que as razões para ambos os grupos serão as mesmas; se este for o caso, então não há DIF. A fórmula é a seguinte:

$$MH = \frac{\sum_j \left(\frac{A_j D_j}{n_{.j}} \right)}{\sum_j \left(\frac{B_j C_j}{n_{.j}} \right)} \qquad (5.18)$$

Dorans e Holland (1992) transformaram esta fórmula para o delta utilizado pela ETS e Swaminathan e Rogers (1990) a transformaram para uma fórmula mais geral dentro da análise da regressão logística, que é uma análise da variância e a qual permite analisar a interação entre grupo e escore no teste, permitindo, inclusive, que o escore seja uma variável contínua e não necessariamente dicotômica.

Métodos da TRI

A TRI produz os parâmetros de dificuldade (b) e de discriminação (a) dos itens para as duas amostras culturalmente diferentes e estes parâmetros podem ser comparados para verificar se são ou não estatisticamente equivalentes ou diferentes. Esta hipótese é testada através do qui-quadrado de Lord (1980; Hulin et al., 1983).

Ao se plotar os índices de dificuldade (b) assim calculados das duas amostras em coordenadas cartesianas, os *b* se alinham ao longo de

uma linha paralela à linha de 45°, que não passa pela origem das coordenadas, mas corta, por exemplo, o eixo dos X, indicando que o teste como um todo é mais difícil para a amostra indicada neste eixo (cf. figura 5-14). Os itens mais difíceis para uma ou outra amostra aparecem endentados em direção ao eixo da amostra para a qual tais itens são particularmente difíceis. Assim, por exemplo, o item #1 é mais difícil para a amostra A, sendo o item #2 mais difícil para a amostra B.

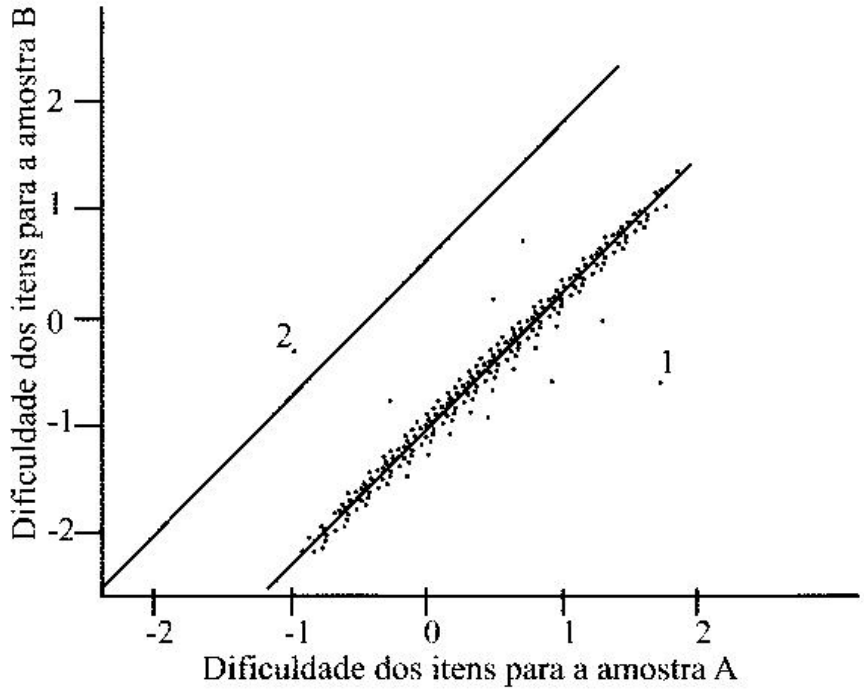


Figura 5-14. Distribuição dos b de duas amostras

b) A *resposta ao acaso*. Os fatores que determinam a resposta ao acaso não são determináveis, pois são, por definição, aleatórios. Tal ocorrência pode ser devida a inúmeros fatores não sistemáticos, como a má disposição do sujeito em responder ao teste, incompreensão das instruções, gozação e outros. A TRI identifica este tipo de resposta através do parâmetro *c*. A Psicometria Clássica trata deste problema através de uma fórmula estatística, a qual supõe que os sujeitos que conhecem a resposta certa assim respondem, ao passo que os que a desconhecem, ou erram ou acertam por acaso. Além disso, supõe que todas as alternativas de resposta sejam equiprováveis, isto é, têm o mesmo valor atrativo de serem selecionadas. Esta última suposição é evidentemente difícil de ser aceita, porque

sobretudo entre as alternativas erradas há praticamente sempre uma ou outra que parece claramente errada (isto é, "está na cara" que está errada!).

De qualquer forma, a fórmula do cálculo do escore no teste, levando em conta as respostas corretas dadas ao acaso, é a seguinte:

$$T_c = A - \frac{E}{n-1} \qquad (5.19)$$

onde,

Tc: Escore corrigido

A: Escore não corrigido (= número de acertos)

E: Número de erros

n: Número de alternativas do item.

Exemplo: O sujeito recebeu escore T de 70 num teste de 100 itens, tendo estes 5 alternativas de resposta, mas apenas uma correta. O sujeito acertou 70 itens, não respondeu 6 e cometeu 24 erros. Qual seria seu escore corrigido por acertos ao acaso?

Resposta: $T_c = 70 - \frac{24}{5-1} = 64$.

c) A *resposta estereotipada*. Trata-se realmente de erros ou de respostas tendenciosas devido a peculiaridades do sujeito que responde, sobretudo ocorrendo em testes de personalidade e de atitude. São devidos a uma estereotipia na resposta. Dois tipos aparecem salientes: a desejabilidade social e as respostas sistemáticas.

c.1) A *desejabilidade social* na verdade corresponde a um traço de personalidade, mas afeta negativamente a objetividade nas respostas de autorrelato. Esta questão foi amplamente discutida por Edwards (1957, 1959; Edwards & Walker, 1961; Edwards, Diers & Walker, 1962; Edwards & Walsh, 1963), que inclusive construiu uma escala para avaliar esta tendência nos sujeitos (Heineman, 1952; Messick & Jackson, 1961). Edwards (1957: vi) define a desejabilidade social como "a tendência dos sujeitos em atribuir a si mesmos, em caso de autodescrição, afirmações de personalidade com valores socialmente desejáveis e em rejeitar aquelas

com valores socialmente indesejáveis". Essa atitude não representa uma vontade de falsear os dados, mas é um desejo (inconsciente) de se apresentar bem diante dos outros. O sujeito não procura intencionalmente mentir sobre si mesmo (neste caso, seria mentira), mas o faz sem dar-se conta disso: quer simplesmente aparecer com bons olhos diante dos outros. Esta tendência é tão comum que parece um traço universal do ser humano. Ela é, igualmente, um problema praticamente sempre presente em inventários de personalidade. A maneira de controlar esta tendência tem sido a elaboração de uma escala de desejabilidade e incluí-la no inventário. Assim, um traço a mais é mensurado pelo inventário; mas fica difícil saber o que fazer com tal dado que alerta sobre o fato de que o sujeito pode bem ter utilizado a mesma tática de resposta aos demais traços medidos pelo inventário. Pelo menos, fica o alerta para a interpretação dos resultados do inventário, quando índices elevados de desejabilidade social estão presentes nos respondentes.

c.2) A *resposta sistemática*, por outro lado, representa falhas de julgamento. Há uma série frustrante deste tipo de falhas (correntemente chamados de erros de resposta): efeito de halo, leniência, tendência central, contraste, proximidade, e outras. O controle destes vieses tem se mostrado ainda bastante falho na utilização de escalas de avaliação.

O *efeito de halo* foi cunhado por Thorndike (1920) e ocorre quando "um avaliador tende a avaliar um indivíduo de modo semelhante sobre todas as dimensões" (Guilford, 1959: 146). Este erro é inversamente proporcional à variância nas respostas (Borman, 1975), acarretando altas correlações entre diferentes fatores (Gillinsky, 1947; Taylor & Hastman, 1956) e reduzidos desvios padrões (Bernardin & Walter, 1977).

O *erro de leniência* consiste em dizer "apenas coisas boas a respeito de todo o mundo" (Dunnette, 1983). Estatisticamente, esta tendência é definida como "uma mudança significativa na média das avaliações na direção favorável, de uma condição de avaliação para outra" (Sharon & Bartlett, 1969: 252).

A *tendência central* ocorre quando um avaliador tende a colocar todos os sujeitos no centro da escala. É uma tendência na qual "avaliadores hesitam proferir julgamentos extremos... e talvez ocorre mais normalmente quando avaliadores não conhecem suficientemente bem os avaliandos" (Guilford, 1954: 278).

Erro de contraste consiste na tendência das pessoas avaliarem os outros ao oposto do que se avaliam a si mesmas. Os outros se tornam o contraponto da autoavaliação. Quem é organizado acha todos os outros desleixados, dizia Murray (1938).

O controle destes e outros tantos erros da resposta se apresenta difícil, dado que eles têm origem na própria personalidade do sujeito que responde, tratando-se, portanto, de outros traços da própria personalidade. Tem-se inventado maneiras de contornar tais vieses, eliminando, por exemplo, o ponto central (neutro) da escala para inviabilizar a tendência central ou eliminar a parte inferior da escala para descaracterizar a leniência, mas tais investidas não têm surtido efeitos suficientes e claros e, assim, estas tendências ainda continuam sendo um problema substancial na medida da personalidade e das atitudes.

Conclusão

Com a entrada da TRI, a análise dos itens em Psicometria se tornou extremamente elaborada. Vamos dar um exemplo que sumariza as informações que a análise dos itens traz para cada item de um teste, informações estas que procedem tanto da TCT quanto da TRI e, até, de outras técnicas estatísticas como a DIF e a análise fatorial. A tabela 5-12 dá essas informações.

Tabela 5-12. Sumário das informações sobre os itens

	Análise gráfica			TRI				TCT				AF	DIF
Item	Dif	Disc	Chu	a	b	c	Res	PC	r_{iT}	$r_{i\theta}$	r_{iC}	Car	Viés
1	6	45	20	0,62	-2,06	0,14	0,64	88	0,29	0,27	0,40	0,46	
2	8	50	20	0,96	-1,62	0,13	1,09	88	0,45	0,49	0,50	0,69	
3	10	45	20	0,61	-0,72	0,14	0,85	71	0,35	0,33	0,38	0,45	
4	16	82	20	1,28	0,11	0,12	1,18	54	0,57	0,56	0,61	0,67	
5	25	75	20	1,01	1,60	0,12	1,48	22	0,31	0,28	0,31	0,35	

Nesta tabela 5-12, as informações sobre os itens são trazidas por:

1) Análise gráfica:

– Dificuldade (Dif): numa escala que corresponde ao escore total no teste (no caso do TNVRI, uma escala que vai de 0 a 30)

– Discriminação (Disc): ângulo da curva, que vai de 0 a 90

– Chute (Chu): percentagem

2) TRI:

– a: índice de dificuldade

– b: índice de discriminação

– c: índice do chute

– Res.: índice de adequação do modelo

3) TCT

– PC (proporção correta, de acertos): índice de dificuldade

– r_{iT} (correlação item-total): índice de discriminação

– $r_{i\theta}$ (correlação item-teta): índice de validade

– r_{iC} (correlação item-critério): índice de validade

4) Análise Fatorial (AF: carga fatorial): índice de validade

5) Análises DIF: índice de viés do item.

Se você tiver todas essas informações sobre cada item de um teste, certamente você produziu uma carteira de identidade para cada item, que lhe permite (1) tomar decisões sobre a qualidade do mesmo e (2) colocá-lo num banco de itens sem que ele perca a sua identidade, porque você tem uma dúzia de indicadores individuais ou entradas nesta carteira de identidade do item, os quais o tornam inconfundível (cf. cap. 10 sobre banco de itens).

CAPÍTULO 6
Validade dos testes

1 – Introdução

A validade constitui um parâmetro da medida tipicamente discutido no contexto das ciências psicossociais. Ela não é corrente em ciências físicas, por exemplo, embora haja nessas ciências ocasiões em que tal parâmetro se aplicaria. Nestas últimas ciências, a preocupação principal na medida se centra na questão da precisão, a dita calibração dos instrumentos. Esta é importante também na medida em ciências psicossociais, mas ela não tem nada a ver, conceitualmente, com a questão da validade. A razão disto está no fato de que a validade diz respeito ao aspecto da medida de ser *congruente* com a propriedade medida dos objetos e não com a exatidão com que a mensuração, que descreve esta propriedade do objeto, é feita. Em Física, o instrumento é um objeto físico que mede propriedades físicas; então parece fácil se ver que a propriedade do objeto mensurante é ou não congruente com a propriedade do objeto medido. Tome, por exemplo, o caso da propriedade "comprimento" do objeto. O instrumento que mede esta propriedade (comprimento), isto é, o metro, usa a sua propriedade de comprimento para medir a comprimento de outro objeto; então estamos medindo comprimento com comprimento, tomados estes termos univocamente. Não há necessidade de provar que a propriedade "comprimento" do metro seja congruente com a mesma propriedade no objeto medido; os termos são unívocos, eles são conceitualmente equivalentes, aliás, idênticos.

O caso já se torna menos claro quando, por exemplo, o astrônomo mede a propriedade "velocidade" galáctica de aproximação ou afastamento via efeito Doppler, onde a aproximação/afastamento das linhas espectrais da luz da galáxia seria o instrumento da medida. Aqui já temos, na verdade, um problema de validade do instrumento de medida, a saber, é verdade ou não que as distâncias das linhas espectrais *têm a ver com* a

velocidade das galáxias? Pode-se fazer tal suposição, mas ela tem que ser demonstrada empiricamente, de alguma maneira, isto é, pelo menos em suas consequências, em hipóteses dela derivadas ou deriváveis e verificáveis. Neste caso específico, o problema da precisão da medida diz respeito à quão exata pode ser feita a mensuração das distâncias entre as linhas espectrais, ao passo que o de validade diz respeito a se esta medida das distâncias das linhas espectrais, por mais exata e perfeita que ela possa ser, tem algo a ver ou não com a velocidade de afastamento da galáxia. Em outras palavras, a validade em tal caso diz respeito à demonstração da adequação (legitimidade) da representação ou da modelagem da velocidade galáctica via distâncias das linhas espectrais.

Este caso da astronomia ilustra o que tipicamente acontece com a medida em ciências psicossociais e, consequentemente, torna a prova da validade dos instrumentos nestas ciências algo fundamental e crucial, isto é, é uma condição *sine qua non* demonstrar a validade dos instrumentos nestas ciências. Isto é particularmente o caso nos enfoques que, em Psicologia, trabalham com o conceito de traço latente, onde se deve demonstrar a correspondência (congruência) entre traço latente e sua representação física (o comportamento). Não causa estranheza, portanto, que o problema de validade tenha tido, na história da Psicologia, uma posição central na teoria da medida, constituindo-se, na verdade, no seu parâmetro fundamental e indispensável. Aliás, a história deste parâmetro é repleta de diatribes que espelham concepções teóricas antagônicas da própria teoria psicológica. À questão de "como legitimar ou justificar a pertinência da medida do comportamento humano?" foram dadas respostas diferentes na história da Psicometria. Podemos ilustrar esta diatribe, distinguindo várias etapas de predominância de uma concepção do parâmetro validade sobre outras e que aparecem sempre atreladas a uma concepção mais geral da própria Psicologia, como já anotava Anastasi em 1986.

Com efeito, poderíamos delinear, em traços bem gerais, a história do parâmetro da validade em três períodos, onde aparece, em cada um deles, a predominância de um dos tipos atualmente conhecidos de validade, desde o famoso trabalho de Cronbach e Meehl (1955), expressos sob o modelo trinitário, a saber, a validade de conteúdo, de critério e de construto.

1º Período: 1900 – 1950: Predomínio da validade de conteúdo

Nesta época estavam em voga as teorias da personalidade e com elas predominava o interesse pelos traços de personalidade (tipos, tem-

peramentos, traços, aptidões, etc.). Estas teorias (psicanálise, fenomenologia, gestaltismo, etc.) apresentavam em geral pouca fundamentação empírica, assumindo um caráter bastante nebuloso, quando não fantasioso. Nesta atmosfera, os testes dos traços eram considerados válidos na medida em que seu conteúdo correspondesse ao conteúdo dos traços teoricamente definidos pela teoria psicológica em questão.

Afora alguns poucos (teste de Binet-Simon, de Raven, de Thurstone e alguns testes projetivos ainda em voga), as dezenas de testes criados nesta época já fazem parte de uma história passada; eles podem ser ditos representantes da pré-história dos testes psicológicos.

2º Período: 1950 – 1970: Predomínio da validade de critério

Prevalecia em Psicologia o enfoque do behaviorismo skinneriano, que influenciou também a Psicometria. Os testes eram concebidos como uma amostra de comportamentos e que tinham como função predizer outros comportamentos ou comportamentos futuros. Este teste era, consequentemente, válido se predizia com precisão os comportamentos numa futura ou outra condição, esta se tornando, assim, o critério de validade do teste. Não interessava saber por que o teste predizia, bastava mostrar que de fato ele o fazia e isto era o critério de sua validade. Este modo de conceber os testes ainda persiste hoje em dia, mas parece que aos poucos sua relevância vai se tornando secundária, tornando-se tão somente uma etapa, juntamente com a validade de conteúdo, no processo de elaboração dos testes psicológicos (Anastasi, 1986).

Este período se caracteriza por uma acentuada fuga do pensar teórico que definia a época anterior. O teste não era mais construído para representar traços de personalidade, mas os itens (tarefas) eram selecionados a partir de um grande elenco (*pool of items*) que parecia se referir àquilo para o qual se queria uma medida, fazendo uso praticamente exclusivo e *a posteriori* de análises estatísticas, especialmente a correlação. Não era mais a teoria psicológica e sim a estatística que definia a qualidade do teste. Este processo de empirismo cego se assemelha ao pescador que lança a rede não importa onde para ver o que pode colher e em cima do colhido decidir o que quer. Neste processo tipicamente se perdem "toneladas" de itens puramente por não satisfazerem critérios estatísticos (Kurtz, 1948; Cureton, 1950; Primoff, 1952). Esta atitude dos psicometristas de então têm suas razões históricas de ser. Eles queriam se desfazer do que

lhes parecia um teorizar gratuito e fantasioso do início do século XX em Psicologia. Contudo, já na década de 1970, os psicometristas procuravam voltar a um teorizar psicológico mais relevante e em cima dele elaborar seus testes, dando início ao terceiro período na concepção dos testes e de sua validade.

3º Período: 1970 – Presente: Predomínio da validade de construto

Este período teve suas fontes históricas no artigo de Cronbach e Meehl (1955) sobre o modelo trinitário da validade (conteúdo, critério, construto). Eles próprios já diziam que a validade de construto exigia um novo tipo de teorizar em Psicometria. Entretanto, o impacto prático desta visão dos autores só se faria sentir após os anos de 1970. Na verdade, a volta à teoria psicológica em Psicometria se deve a vários fatores, salientando-se:

1) Preocupação com desenvolver a teoria da personalidade e inteligência em especial, com maior base empírica, valendo-se sobretudo das técnicas da análise fatorial (Comrey, 1970: Guilford, 1967; Jackson, 1974; Millon, 1983; Cattell, 1965; Cattell & Stice, 1957; Cattell & Warburton, 1967);

2) Estudos dos processos cognitivos (Sternberg, 1977, 1984; Sternberg & Detterman, 1986; Sternberg & Rifkin, 1979);

3) Estudos do processamento da informação (Newell, Shaw & Simon, 1958a; Newell, Shaw & Simon, 1958b);

4) Insatisfação com os resultados decepcionantes do uso dos testes na situação educacional e do trabalho. Na clínica se utilizavam ainda bastante os testes projetivos, onde predominava, aliás, ainda o pensamento da primeira época dos testes baseados nas teorias dos traços de personalidade;

5) O impacto da Teoria de Resposta ao Item (TRI) com sua insistência no traço latente. A influência decisiva desta teoria ocorre somente após os anos de 1980, retardo devido ao atraso na área da informática para fazer uso prático das análises estatística complexas que tal enfoque exige.

A preocupação agora, na validação dos instrumentos psicológicos, se concentra na validade de construto ou dos traços latentes. Não está ainda finalizada a disputa entre a ênfase ou nos traços ou nas situações

(*construct-centered* vs. *task-centered*) ou, como diz Messick (1994), entre a avaliação *task-driven* versus *construct-driven*. Parece, entretanto, que o conceito de validade dos testes psicológicos irá finalmente se reduzir à validade de construto, sendo o de conteúdo e o de critério apenas aspectos da validade de construto (Anastasi, 1986; Messick, 1989, 1994; Embretson, 1983; Wiggins, 1989; Cronbach, 1989, o qual já em 1955, de algum modo, previa tal desenlace). Esta tendência é obviamente favorecida também pelos psicólogos da linha cognitivista (Sternberg, 1985, 1990; Gardner, 1983).

Vamos voltar, agora, para o tema da validade dos testes. Nos manuais de Psicometria costuma-se definir a validade de um teste dizendo que ele é válido se de fato mede o que supostamente deve medir. Embora esta definição pareça uma tautologia, na verdade ela não é, considerada a teoria psicométrica exposta neste trabalho sobre o traço latente. O que se quer dizer com esta definição é que, ao se medirem os comportamentos (itens), que são a representação do traço latente, está-se medindo o próprio traço latente. Tal suposição é justificada se a representação comportamental for legítima. Esta legitimação somente é possível se existir uma teoria prévia do traço que fundamente que a tal representação comportamental constitui uma hipótese dedutível desta teoria. A validade do teste (este constituindo a hipótese), então, será estabelecida pela testagem empírica da verificação da hipótese. Pelo menos, esta é a metodologia científica. Assim, fica muito estranha a prática corrente na Psicometria de se agrupar intuitivamente uma série de itens e, a posteriori, verificar estatisticamente o que eles estão medindo. A ênfase na formulação da teoria sobre os traços foi muito fraca no passado; com a influência da Psicologia Cognitiva esta ênfase felizmente está voltando ou deverá voltar ao seu devido lugar na Psicometria.

Aliás, a Psicometria clássica entende por "aquilo que supostamente deve medir" como sendo o "critério", este representado por teste paralelo. Assim, este "aquilo que" é o *traço latente* na concepção cognitivista da Psicometria e é o *critério* (escore no teste paralelo) na visão comportamentalista.

Diz Anastasi (1986: 3) que o processo de validação de um teste "inicia com a formulação de definições detalhadas do traço ou construto, derivadas da teoria psicológica, pesquisa anterior, ou observação sistemática e análises do domínio relevante do comportamento. Os itens do

teste são então preparados para se adequarem às definições do construto. Análises empíricas dos itens seguem, selecionando-se finalmente os itens mais eficazes (i.é, válidos) da amostra inicial de itens"[1].

A validação da representação comportamental do traço, isto é, do teste, embora constitua o ponto nevrálgico da Psicometria, apresenta dificuldades importantes que se situam em três níveis ou momentos do processo de elaboração do instrumento, a saber, ao nível da teoria, da coleta empírica da informação e da própria análise estatística da informação.

No nível da teoria se concentram talvez as maiores dificuldades. Na verdade, a teoria psicológica se encontra ainda em estado embrionário, destituída quase que totalmente de qualquer nível de axiomatização, resultando disto uma pletora de teorias, muitas vezes até contraditórias. Basta lembrar de teorias como behaviorismo, psicanálise, psicologia existencialista, psicologia dialética e outras, que, existindo simultaneamente, postulam princípios irredutíveis entre as várias teorias e pouco concatenados dentro de uma mesma teoria ou, então, em número insuficiente para se poder deduzir hipóteses úteis para o conhecimento psicológico. Havendo esta confusão no campo teórico dos construtos, torna-se extremamente difícil para o psicometrista operacionalizar estes mesmos construtos, isto é, formular hipóteses claras e precisas para testar ou, então, formular hipóteses psicologicamente úteis. Ainda quando a operacionalização for um sucesso, a coleta da informação empírica não será isenta de dificuldades, como, por exemplo, a definição inequívoca de grupos critérios onde estes construtos possam ser idealmente estudados. Mesmo ao nível das análises estatísticas encontramos problemas. Pela lógica da elaboração do instrumento, a verificação da hipótese da legitimidade da representação dos construtos se faz por análises do tipo da análise fatorial (confirmatória), que procura identificar, nos dados empíricos, os construtos previamente operacionalizados no instrumento. Mas, acontece que a análise fatorial faz algumas postulações fortes que nem sempre se coadunam com a realidade dos fatos. Por exemplo, a análise fatorial assume que as respostas dos sujeitos aos itens do instrumento são determinadas por uma relação linear destes com os traços latentes. Há, ainda, o grave problema da rotação dos

---

1. A questão da elaboração de testes psicológicos é detalhadamente tratada no livro organizado por este autor, *Instrumentos Psicológicos: Manual Prático de Elaboração*. Brasília, DF: LabPAM/IBAPP, 1999.

eixos, a qual permite a demonstração de um número sem fim de fatores para o mesmo instrumento (cf. capítulo 11 sobre Análise Fatorial)[2].

Diante destas dificuldades, os psicometristas recorrem a uma série de técnicas para viabilizar a demonstração da validade dos seus instrumentos. Fundamentalmente, estas técnicas podem ser reduzidas a três grandes classes (o modelo trinitário): técnicas que visam a validade de construto, validade de conteúdo e validade de critério (APA, 1954).

2 – Validade de construto

A validade de construto ou de conceito é considerada a forma mais fundamental de validade dos instrumentos psicológicos e com toda a razão, dado que ela constitui a maneira direta de verificar a hipótese da legitimidade da representação comportamental dos traços latentes e, portanto, se coaduna exatamente com a teoria psicométrica aqui defendida. Historicamente, o conceito de construto entrou na Psicometria através da *American Psychological Association Committee on Psychological Tests* que trabalhou entre 1950 e 1954 e cujos resultados se tornaram as recomendações técnicas para os testes psicológicos (APA, 1954).

O conceito de validade de construto foi elaborado com o já clássico artigo de Cronbach e Meehl (1955) *Construct validity in psychological tests*, embora o conceito já tivesse uma história sob outros nomes, tais como validade intrínseca, validade fatorial e até validade aparente (*face validity*). Estas várias terminologias demonstram a confusa noção que construto possuía. Embora tenham tentado clarear o conceito de validade de construto, Cronbach e Meehl ainda o definem como a característica de um teste enquanto mensuração de um atributo ou qualidade, o qual não tenha sido "*definido operacionalmente*". Reconhecem, entretanto, que a validade de construto reclamava por um novo enfoque científico. De fato, definir esta validade do modo que eles a definiram parece um pouco estranho em ciência, dado que conceitos não definidos operacionalmente não são suscetíveis de conhecimento científico. Conceitos ou construtos são cientificamente pesquisáveis somente se forem, pelo menos, passíveis de representação comportamental adequada. Do contrário, serão conceitos

2. Veja capítulo sobre análise fatorial neste livro; também o livro do mesmo autor "Análise fatorial para pesquisadores". Petrópolis: Vozes, s.d.

metafísicos e não científicos. O problema está em que, sintetizando a atitude geral dos psicometristas da época, para definir validade de construto, os autores partiram do teste, isto é, da representação comportamental, em vez de partir da teoria psicométrica que se fundamenta na elaboração da teoria do construto (dos traços latentes). O problema não é descobrir o construto a partir de uma representação existente (teste), mas sim descobrir se a representação (teste) constitui uma representação legítima, adequada do construto. Este enfoque exige uma colaboração, bem mais estreita do que existe, entre psicometristas e Psicologia Cognitiva.

A validade de construto de um teste pode ser trabalhada sob vários ângulos: a análise da representação comportamental do construto, a análise por hipótese, a curva de informação da TRI, além do falsete estatístico do erro de estimação da TCT. Vamos iniciar com este último.

2.1 – O erro de estimação

Esta forma de avaliar a validade de um teste era típico da psicometria clássica. Como vimos nos capítulos 3 e 4, este modelo de Psicometria é um modelo que poderíamos chamar de positivista, uma vez que ele se fundamenta exclusivamente nos dados empíricos coletados de um conjunto de itens agrupados inicialmente mais ou menos de maneira intuitiva. Na verdade, o teste (conjunto de itens) é construído através da seleção de uma amostra de itens coletados de um universo de itens que *parecem* medir um dado construto. Esta maneira de construir instrumentos psicométricos se fundamenta na ideia de que existe, para cada construto, um universo indefinido de itens (*pool of itens*), do qual uma amostra é extraída para constituir o teste. Como é que se sabe inicialmente que os itens incluídos na amostra se referem a um construto somente ou que estamos retirando itens de um universo unidimensional para compor o teste? Apela-se aqui à famosa ou malfadada validade aparente (*face validity*), isto é, os itens *parecem* estar se referindo à mesma coisa! Por mais estranho que isto pareça, honestamente é o que se faz nesta tradição positivista da Psicometria. É que nesta tradição falta todo o teorizar prévio sobre o construto (traço latente) para o qual se quer construir o instrumento de medida. Sem os procedimentos teóricos sobre o traço latente, os itens não são construídos para representá-lo comportamentalmente, mas são coletados mais ou menos a esmo ("chutados") com base na validade aparente e verificados depois, através de análises estatísticas, para ver se de fato eles

estão ou não se referindo a alguma coisa (construto) comum. Assim, a Psicometria se torna, no máximo, um ramo da Estatística, como, aliás, era normalmente definida, e não um ramo da Psicologia, como deve ser concebida. Para a Estatística, o número é número, não interessa donde ele vem; mas para a Psicologia (Psicometria) o número é uma representação de conteúdo psicológico, então interessa muito donde este número vem. Na tradição clássica da Psicometria apela-se demasiadamente à Estatística para salvar a teoria psicológica. Isso não dá. Não se pode abdicar da teoria psicológica em favor da Estatística. É preciso, primeiramente, fazer e avançar a teoria psicológica (dos traços latentes) e apelar, em seguida, à Estatística para auxiliar a tomar decisões mais objetivas sobre a demonstração de hipóteses psicologicamente significativas e relevantes, estas deduzidas da teoria psicológica e não levantadas intuitiva e aleatoriamente. A Psicometria, clássica e também a moderna, estão a necessitar urgentemente da ajuda da Psicologia Cognitiva neste particular, que a possa instrumentar com a teoria dos traços latentes para os quais ela quer desenvolver instrumentos de observação quantitativa (medida).

De qualquer forma, também na TCT se procura demonstrar a validade dos testes. Como é que era isto feito? Vejamos.

Neste contexto, a Psicometria clássica procura legitimar a validade de um instrumento dentro do conceito de *erro de estimação*, isto é, quanto o escore obtido pelo sujeito no teste se afasta do escore verdadeiro.

A fórmula para o cálculo do erro de estimação, na qual um critério é predito a partir de um teste, é a seguinte:

$$EE = s_C \sqrt{1 - r_{TC}^2} \qquad (6.1)$$

onde,

s_C é o desvio padrão da medida do critério

r_{TC}^2 é o coeficiente de validade, isto é, a correlação entre o teste e o critério.

Esta fórmula está baseada na ideia de se computar o erro mínimo que se pode cometer ao se predizer o escore de um teste a partir do escore de um teste paralelo. Para tanto se usa a regressão dos quadrados mínimos, o que dá

$$d = t_1 - r_{12}\left(\frac{s_1}{s_2}\right)t_2 \qquad (6.2)$$

isto é, para predizer o escore num teste paralelo (T_2) a partir de um teste T_1, calcula-se a diferença entre estes dois escores, levando em conta a correlação que o teste T_1 tem com o teste paralelo. Por isso que $d = t_1 - r_{12}t_2$.

Entretanto, como $s_1 = s_2$, já que se trata de testes paralelos, o termo entre parênteses é igual a 1. Assim, a fórmula se reduz a

$$d = t_1 - r_{12}t_2. \qquad (6.3)$$

Para se tomarem os desvios médios (e não os desvios individuais), temos que, primeiro, elevar os desvios (d) ao quadrado, dado que a soma dos desvios em torno da média vai dar sempre 0. Em seguida, somar estes desvios quadrados e dividir tudo por N (número de sujeitos da amostra sobre os quais a soma dos desvios foi feita). Tal procedimento resultará em

$$\frac{\sum d^2}{N} = \frac{\sum (t_1 - r_{12}t_2)^2}{N}. \qquad (6.4)$$

Expandindo, a equação se torna

$$\frac{\sum d^2}{N} = \frac{\sum \left(t_1^2 - r_{12}t_1t_2 - r_{12}t_1t_2 + r_{12}^2t_2^2\right)}{N}$$

$$= \frac{\sum \left(t_1^2 - 2r_{12}t_1t_2 + r_{12}^2t_2^2\right)}{N}$$

$$= \frac{\sum t_1^2}{N} + \frac{r_{12}^2 \sum t_2^2}{N} - \frac{2r_{12}\sum t_1t_2}{N}$$

Como se trata de desvios médios (variância e covariância), a fórmula fica

$$s_d^2 = s_1^2 + r_{12}^2 s_2^2 - 2r_{12}s_1s_2 \qquad (6.5)$$

Como em testes paralelos as variâncias são iguais ($s_1 = s_2$), então $s_1s_2 = s_1^2$ e, assim, a fórmula 6.5 se torna

$$s_d^2 = s_1^2 + r_{12}^2 s_2^2 - 2r_{12}s_1^2 \quad \text{ou}$$

$$\begin{aligned} s_d^2 &= s_1^2 + r_{12}^2 s_1^2 - 2r_{12}^2 s_1^2 \\ &= s_1^2 - r_{12}^2 s_1^2 \\ &= s_1^2\left(1 - r_{12}^2\right) \end{aligned}$$

e extraindo a raiz quadrada, a fórmula finalmente se torna

$$s_d = s_1\sqrt{1 - r_{12}^2} \tag{6.6}$$

onde o subscrito 1 se refere ao teste (teste 1 ou T na nossa fórmula inicial 6.1) e o 2 ao critério (teste 2 ou C na fórmula inicial).

Vê-se pela fórmula que, para se poder obter o erro de estimação, é necessário se possuir a medida de um critério, este supostamente sendo a medida da aptidão. Por mais precário que tal procedimento pareça, é dos poucos de que dispõe a Psicometria Clássica para estabelecer o erro de estimação e, por consequência, a validade de um teste, entendida como a precisão com a qual o teste pode predizer o escore verdadeiro. Esta fórmula deixa claro que se o coeficiente de validade r_{TC}^2 for zero, então o erro de estimação (EE) é igual ao desvio padrão da medida, pois o fator sob a raiz equivaleria a 1. Tal ocorrência implicaria que o teste não é capaz de predizer o escore verdadeiro melhor do que uma simples adivinhação, isto é, ele é totalmente inútil para predizer qualquer coisa. Agora, se o coeficiente de validade for diferente de zero, então o teste tem poder maior de predizer do que uma simples adivinhação. Quanto maior? Vejamos: se o coeficiente de validade fosse igual a 1, o erro de estimação seria zero, pois ele seria o desvio padrão vezes 0; mas suponhamos que um teste tenha coeficiente de validade de 0,80, o que constitui um coeficiente de grandeza extraordinária em termos práticos. Neste caso, qual seria a força de predição do teste com respeito ao critério que pretende medir? Calculando o erro de estimação do teste, temos

$$EE = 1 \times \sqrt{1 - 0{,}80^2} = \sqrt{1 - 0{,}64} = \sqrt{0{,}36} = 0{,}60.$$

Assim, a predição do teste é 40% (1,00 – 0,60) superior à predição feita ao acaso ou por adivinhação. Isso não parece grande coisa dado um coeficiente de validade tão elevado, mas sempre é melhor que a pura adivinhação. Também, felizmente, o erro de estimação e o coeficiente de validade não são os únicos nem os melhores procedimentos para estabelecer a validade de um teste, como veremos a seguir.

Na verdade, há neste procedimento do erro de estimação um certo equívoco ao se supor que o escore verdadeiro seja a medida daquilo que o teste pretende medir. De fato, o escore verdadeiro constitui um agregado de medida daquilo que o teste pretende medir mais características peculiares dos itens que compõem o teste, sem que estas tenham a ver com o que o teste pretende medir. Vejamos:

Se decompusermos a variância que um item produz numa amostra de sujeitos, podemos visualizar a situação apresentada na figura 6-1, como fica claro no contexto da análise fatorial[3]. Segundo esta ilustração, a variância tem três componentes: a variância comum (h^2), a variância específica (s^2) e a variância erro (e^2).

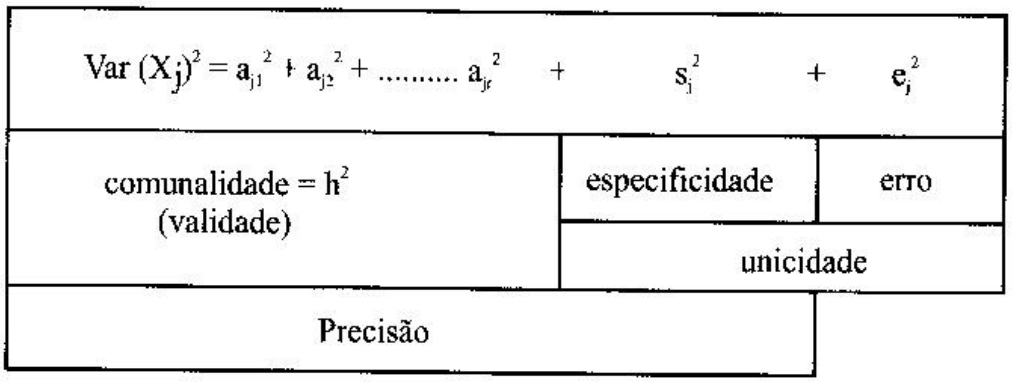


Figura 6-1. Distribuição dos componentes da variância

A variância que tem a ver diretamente com o conceito de validade é a variância comum exclusivamente, pois ela representa a saturação do item no traço latente (ou aptidão), isto é, ela é a covariância, a correlação do item com o traço latente; vale dizer, ela constitui o que o item tem a ver com o traço latente. O restante da variância do item não tem nada a ver com o traço latente e, consequentemente, nada tem a ver com o conceito de validade. Acontece, porém, que o conceito de escore verdadeiro (V) da

3. Cf. nota 2.

Psicometria Clássica inclui, além da variância comum, a variância específica, dado que o outro componente da variância assumida no modelo (T = V + E) é somente o erro. Assim, a variância que serve de base para o cálculo da validade vai ser a mesma que servirá para calcular o coeficiente de precisão, tornando estes dois conceitos (validade e precisão) conceitualmente confusos. Que a variância comum entre na concepção e cálculo da precisão é legítimo, mas não é legítimo que entre na concepção e, consequentemente, no cálculo de validade a variância específica dos itens, a não ser, evidentemente, como é o caso da Psicometria Clássica, que o critério de validade seja outro escore verdadeiro, isto é, outro teste (paralelo). Então, vê-se que o critério de validade não é o traço latente e sim o escore verdadeiro de outro teste paralelo. É no que dá a falta da teoria psicológica como base para fundamentar estes conceitos paramétricos, que se tornam, desta forma, conceitos puramente estatísticos.

2.2 – Análise da representação

São utilizadas duas técnicas como demonstração da adequação da representação do construto pelo teste: a análise fatorial e a análise da consistência interna.

a) Consistência interna

A análise da consistência interna consiste em calcular a correlação que existe entre cada item do teste e o restante dos itens ou o total (escore total) dos itens. Dado que o item sendo analisado contribui para o escore total, ele teoricamente não deve entrar neste escore, já que é ele que está sendo escrutinado. Assim, a correlação legítima será a do item com o restante dos itens. Esta preocupação é importante quando o número de itens do teste for pequeno, pois neste caso o próprio item em análise afeta substancialmente o escore total a seu favor. Por exemplo, num teste com 10 itens, cada item contribui e influencia o escore total em 10%. Quanto maior, contudo, o número de itens que compõem o teste, a influência de cada item em particular no escore total vai se tornando menos relevante. Em um teste com 100 itens, por exemplo, cada item afeta o escore total em apenas 1%. Consequentemente, no caso de teste com grande número de itens ($n \geq 30$), a correlação do item com o escore total ou com o restante dos itens não vai fazer diferença relevante.

Um exemplo para cálculo desta correlação segue na tabela 6-1, onde 10 sujeitos responderam um teste de 10 itens numa escala de 5 pontos, obtendo os resultados apresentados na tabela com referência ao item 1 e o restante do teste. Pergunta-se qual a consistência deste item no teste? A resposta é dada pela correlação, no caso sendo r = 0,68, que é muito elevada.

Tabela 6-1. Cálculo da consistência interna de um item em teste de 10 itens

Sujeitos	Item 1 (X)	Restante (Y)	x	y	x^2	y^2	xy
1	5	45	1,5	10,5	2,25	110,25	15,75
2	4	40	0,5	5,5	0,25	30,25	2,75
3	3	35	-0,5	0,5	0,25	0,25	-0,25
4	2	30	-1,5	-4,5	2,25	20,25	6,75
5	4	40	0,5	5,5	0,25	30,25	2,75
6	3	40	-0,5	5,5	0,25	30,25	-2,75
7	2	20	-1,5	-14,5	2,25	210,25	21,75
8	4	30	0,5	-4,5	0,25	20,25	-2,25
9	5	35	1,5	0,5	2,25	0,25	0,75
10	3	30	-0,5	-4,5	0,25	20,25	2,25
Soma	35	345	0	0	10,50	472,50	47,50
Média	3,5	34,5					

$$s_X = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\frac{10,50}{10}} = 1,02 \qquad s_Y = \sqrt{\frac{\sum y^2}{N}} = \sqrt{\frac{472,50}{10}} = 6,87$$

$$r_{XY} = \frac{\sum xy}{N s_X s_Y} = \frac{47,50}{10 \times 1,02 \times 6,87} = 0,68$$

A análise da consistência interna do teste implica no cálculo das correlações de cada item individualmente com o restante do teste. Esta análise apresenta um problema lógico que se situa no escore total. Na verdade, o escore total é o critério contra o qual cada item é avaliado; mas acontece que os itens são os que vão constituir o escore total, antes mesmo de se saber se eles são itens válidos e somáveis (unidimensionais, isto é, que estão medindo um e o mesmo traço latente). O escore total constitui, assim, uma dificuldade, dado que ele somente faz sentido se o teste já é *a priori* homogêneo. De sorte que a correlação de cada item com o escore total já pressupõe que os itens são somáveis, isto é, homogêneos e válidos; em outras palavras, se pressupõe que todos os itens constituam uma representação adequada do traço e de um mesmo traço latente (unidimensionalidade). Além disso, a consistência interna implica em que os itens estejam intercorrelacionados, isto é, que as correlações entre eles mesmos sejam elevadas. Entretanto, as intercorrelações entre os itens não são uma demonstração de que estes estejam medindo um e mesmo construto. Suponha a situação de três itens saturados em três fatores, como segue:

Item	F1	F2	F3
1	0,80	0,30	0,30
2	0,30	0,80	0,30
3	0,30	0,30	0,80

As correlações entre os três itens são todas de 0,57, altas e significativas, mas nem por isso se pode dizer que os três itens estejam medindo uma e a mesma coisa. Na verdade, o item 1 mede especificamente o fator 1, pois está altamente saturado somente neste fator e não nos outros dois, e os outros itens medem outros fatores. Consequentemente, a análise da consistência interna dos itens não parece garantir que eles sejam uma representação unidimensional de um construto.

A conclusão que se impõe destas observações é a de que a análise da consistência interna não constitui prova cabal de validade de construto do teste.

b) Análise fatorial[4]

Por outro lado, a análise fatorial tem como lógica precisamente verificar quantos construtos comuns são necessários para explicar as covariâncias (as intercorrelações) dos itens. As correlações entre os itens são explicadas, pela análise fatorial, como resultantes de variáveis-fonte que seriam as causas destas covariâncias. Estas variáveis-fonte são os construtos ou traços latentes de que fala a Psicometria. A análise fatorial também postula que um número menor de traços latentes (variáveis-fonte) é suficiente para explicar um número maior de variáveis observadas (itens), como se verifica na figura 6-2.

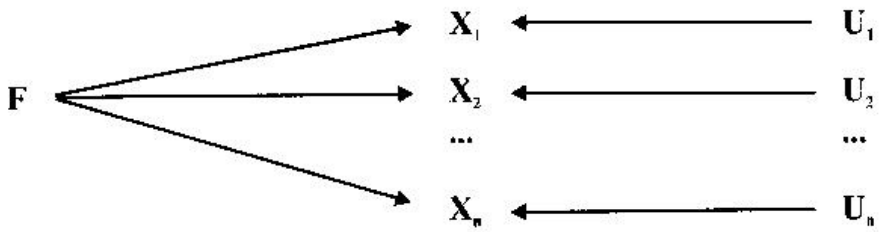


Figura 6-2. Representação do modelo fatorial

O modelo da figura 6-2 mostra que *n* variáveis (X) podem ser explicadas por um fator comum a todas as variáveis (F) mais um fator específico para cada uma delas (U). De sorte que cada variável tem sua equação expressa em termos destes dois fatores. Por exemplo:

$X_1 = a_1F + d_1U_1$

O a_1 é a saturação, a correlação, a covariância (dita carga fatorial) da variável X_1 no fator F. Ela representa o percentual de relação que ela tem com o fator, isto é, quanto porcento ela se constitui em representação do fator (traço latente); indica, em outras palavras, se ela é uma boa representação comportamental do traço latente. Além disso, as cargas fatoriais são as que determinam a correlação entre as próprias variáveis empíricas. Assim, a correlação entre X_1 e X_2 é definida por a_1a_2[5].

---

4. Cf. nota 2.

5. Cf. nota 2.

Desta forma, a validade de construto de um teste é determinada pela grandeza das cargas fatoriais (que são correlações que vão de -1 a +1) das variáveis no fator, sendo aquelas a representação comportamental deste fator, que, por sua vez, é o traço latente para o qual elas foram inicialmente elaboradas como representação empírica. Estas cargas fatoriais representam a parte fundamental do escore verdadeiro (V) da equação da Psicometria Clássica: T = V + E. Dizemos parte fundamental, porque outra parte do V é constituída pela contribuição específica do item (contida no fator U do modelo fatorial) para o escore empírico T do teste. De fato, a variância total de um item ou variável pode ser decomposta em variância comum, variância específica e variância erro, como vimos acima.

A variância comum representa o que as variáveis do teste têm em comum (expressa pelas intercorrelações entre elas) e que é recolhida nas cargas fatoriais no fator comum F e é esta que constitui a questão da validade do teste, isto é, quanto do traço latente (fator F) é representado empiricamente pelas variáveis (itens). O restante da variância dos itens é recolhida na chamada unicidade (U) de cada item que representa tanto o que é específico de cada um deles quanto os erros de medida. Estes dois últimos aspectos da variância (especificidade e erro) são agrupados num conceito só, a saber, a unicidade, porque eles não contribuem para a validade do teste, pois é a porção do item que não constitui representação do traço latente.

Se não houvesse dificuldades com o modelo da análise fatorial, esta constituiria uma demonstração empírica cabal da validade de construto de um teste, pois forneceria a expressão exata de quanto o teste estaria representando o traço latente. Mas, infelizmente, a análise fatorial apresenta alguns problemas importantes. Duas razões são a preocupação principal neste particular. Primeiramente, o modelo fatorial se fundamenta em equações exclusivamente lineares entre variáveis e fatores. Embora seja rotineiro em matemática tentar, em primeira aproximação, um modelo linear, parece difícil se admitir que as intercorrelações empíricas entre os itens e a relação destes com os fatores (variáveis-fonte) possam ser todas reduzidas a equações lineares. Isto é tanto mais plausível, quando se observa que em quiçá nenhum campo da Psicologia e ciências psicossociais em geral se encontram tais equações. Encontram-se, sim, equações logarítmicas, exponenciais e outras, isto é, equações não lineares, como, por exemplo, nas leis da psicofísica (leis de potência) e da análise experimen-

tal do comportamento (lei da igualação). Em segundo lugar, existe o grave problema da rotação dos eixos, para a qual não existe nenhum critério objetivo, a não ser a interpretabilidade psicológica (semântica) dos fatores. Esta ocorrência permite, em tese, a descoberta de qualquer fator que se queira, tornando a solução extremamente arbitrária. Contudo, se o teste foi construído via teoria psicológica de traços latentes e não a esmo (como coleta de uma amostra de itens a partir de um universo arbitrário deles, como é praxe corrente na construção de testes), temos ali um critério objetivo de rotação dos eixos em função dos traços latentes para os quais os itens foram inicialmente construídos como representação comportamental. Neste caso, a análise fatorial será utilizada como teste de hipótese e não como pesca de hipóteses, assumindo, assim, a Estatística, como é legítimo, o papel de testagem de hipóteses psicológicas formuladas pela teoria psicológica e não o papel de criar ela (Estatística) as hipóteses psicológicas (*a posteriori*).

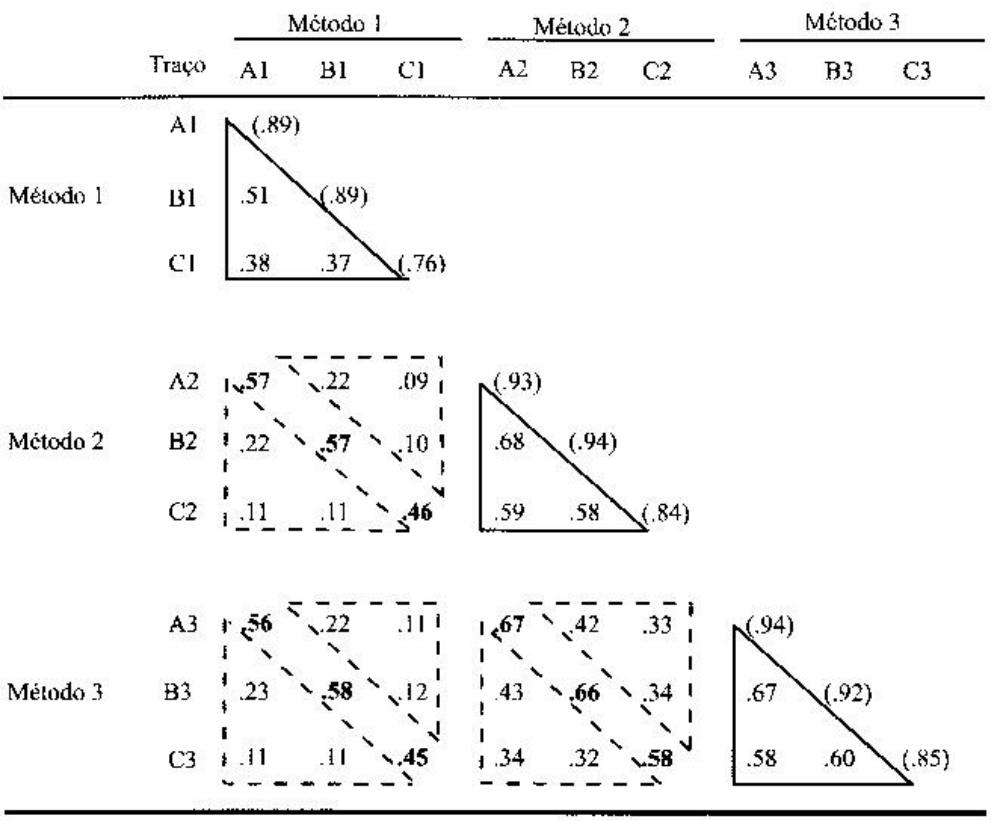
2.3 – Análise por hipótese

Esta análise se fundamenta no poder de um teste psicológico ser capaz de discriminar ou predizer um critério externo a ele mesmo; por exemplo, discriminar grupos-critério que difiram especificamente no traço que o teste mede. Este critério é procurado de várias formas, havendo quatro entre as mais salientes e normalmente utilizadas, a saber, a validação convergente-discriminante, idade, outros testes do mesmo construto e a experimentação.

A técnica da *validação convergente-discriminante* (Campbell & Fiske, 1967) parte do princípio de que para demonstrar a validade de construto de um teste é preciso determinar duas coisas: (1) o teste deve correlacionar significativamente com outras variáveis com as quais o construto medido pelo teste deveria, pela teoria, estar relacionado (validade convergente) e, (2) não se correlacionar com variáveis com as quais ele teoricamente deveria diferir (validade discriminante).

Campbell e Fiske (1967: 125) apresentam o exemplo da tabela 6-2.

Tabela 6-2. Matriz sintética de Multitraço-Multimétodo (Campbell & Fiske, 1967)



A ilustração apresenta seis blocos de resultados: três triângulos (com linhas inteiras) e três retângulos (com triângulos de linhas pontilhadas). As diagonais dos blocos-retângulo representam as correlações entre as variáveis medidas por diferentes métodos e contêm a diagonal da validade (validade convergente): estes valores devem ser altos para mostrar validade de construto. Os valores fora destas diagonais nestes mesmos blocos (os triângulos de linha pontilhada) representam as correlações entre diferentes variáveis medidas por diferentes métodos: estes valores devem ser pequenos para mostrar validade de construto (validade discriminante). O mesmo deve ocorrer com as correlações fora das diagonais nos blocos-triângulo (com linhas inteiras), que representam os coeficientes entre variáveis diferentes medidas pelo mesmo método (nas diagonais estão os coeficientes de precisão). No caso específico dos dados da tabela, os resultados nos quadrados satisfazem os critérios propostos, mas as correlações fora das diagonais nos triângulos são demasiadamente elevadas para

satisfazerem os critérios de validade discriminante. Há, ao que parece, ali um efeito espúrio do instrumento, a saber, o mesmo teste medindo variáveis diferentes produz correlação entre elas por simples efeito de contiguidade.

Este método funciona se os métodos e as variáveis diferem o suficiente (maximamente) entre si.

A *idade* é utilizada 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 hipótese a ser testada neste método é a de que o teste que mede o traço X, o qual muda claramente com a idade, é capaz de discriminar distintamente grupos de idades diferentes.

A prova que se faz neste caso é a da diferença entre a média no teste de sujeitos mais jovens ($\overline{T}_j$) e a média de sujeitos mais adultos ($\overline{T}_a$), a saber

$$t = \frac{\overline{T}_a - \overline{T}_j}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_j^2}{n_j}}} \qquad (6.7)$$

onde,

$\overline{T}_j$ e $\overline{T}_a$ são as médias no teste do grupo jovem e do grupo adulto

s_j^2 e s_a^2 são as variâncias destas médias

n_j e n_a é o número de sujeitos nos dois respectivos grupos.

Os graus de liberdade para verificar a significância do teste de Student "t" são $n_j + n_a - 2$.

Na história dos testes psicológicos, este procedimento de validação foi talvez o primeiro a ser utilizado quando Binet e Simon (1908) utilizaram o critério de diferenciação por idade na seleção dos itens do seu famoso teste de inteligência. Embora a preocupação explícita dos autores

fosse construir um teste que fosse capaz de predizer o desempenho acadêmico de alunos do primeiro grau, eles se basearam numa hipótese de caráter conceitual, isto é, de que as habilidades cognitivas aumentam sistematicamente com a idade cronológica (na infância) e, para medi-las, escolheram tarefas específicas cuja execução correta correspondia a determinada faixa etária.

Este método contém um problema, o qual consiste no fato de que a maturação psicológica pode assumir dimensões e conotações muito distintas em culturas diferentes, por um lado; por outro, outras variáveis que não o traço em questão podem estar dependentes desta maturação, dificultando ou impossibilitando a definição dos grupos-critério somente em função da idade. Assim, se outras variáveis variam com a idade, pode bem ser que estas sejam as responsáveis pelas mudanças no escore e não a idade especificamente. Isto não seria grave problema se estas outras variáveis covariassem sistematicamente com o traço latente que o teste quer medir e, além disso, variassem do mesmo modo em qualquer contexto cultural ou socioeconômico, o que obviamente é difícil de assumir. Dentro de uma mesma cultura, o método pode se apresentar como importante para a determinação da validade de construto.

A *correlação com outros testes* que meçam o mesmo traço é também utilizada como demonstração da validade de construto. O argumento é de que, se um teste X mede validamente o traço Z e o novo teste N se correlaciona altamente com o teste X, então o novo teste mede o mêsmo traço medido por aquele teste. Veja exemplo na tabela 6-3, onde $n = N - \overline{N}$ e $x = X - \overline{X}$.

Tabela 6-3. Cálculo da correlação entre teste N e teste X

Sujeitos	N	X	n	x	n^2	x^2	nx
1	50	45	12,7	8,7	161,29	75,69	110,49
2	45	30	7,7	-6,3	59,29	39,69	-48,51
3	30	20	-7,3	-16,3	53,29	265,69	118,99
4	25	30	-12,3	-6,3	151,29	39,69	77,49
5	43	50	5,7	13,7	32,49	187,69	78,09
6	20	20	-17,3	-16,3	299,29	265,69	281,99
7	38	36	0,7	-0,3	0,49	0,09	-0,21
8	47	50	9,7	13,7	94,09	187,69	132,89
9	35	37	-2,3	0,7	5,29	0,49	-1,61
10	40	45	2,7	8,7	7,29	75,69	23,49
Soma	373	363	0	0	864,10	1138,10	776,10
Média	37,3	36,3					

$$s_N = \sqrt{\frac{\sum n^2}{N}} = \sqrt{\frac{864,1}{10}} = 9,30$$

$$s_X = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\frac{1138,1}{10}} = 10,67$$

$$r_{NX} = \frac{\sum nx}{N \times s_N \times s_X} = \frac{773,1}{10 \times 9,30 \times 10,67} = 0,78$$

Esta técnica também contém um problema, o qual consiste no fato de que normalmente um teste de um traço qualquer não se apresenta com tal pureza a se poder afirmar que ele mede exclusivamente o tal traço. De fato, ele mede o traço em termos de um certo nível de covariância: por exemplo, existe uma correlação de 0,70 entre o teste X e o traço, o que equivale a uma comunalidade de 49% entre os dois. Agora, o novo teste N correlaciona 0,78 (exemplo da Tabela 6-3) com aquele teste X, havendo, portanto, comunalidade de 61% entre os dois testes. Qual será, neste caso, a comunalidade do novo teste com o traço em si? Por azar poderia aconte-

cer que a comunalidade de 61% entre os dois testes ocorra precisamente com os 51% do primeiro teste que não covariam com o traço; neste caso, a comunalidade do novo teste com o traço seria de apenas 10%, isto é, o novo teste seria uma representação quase totalmente equivocada do traço, como ilustra a figura 6-3, na faixa duplamente rajada.

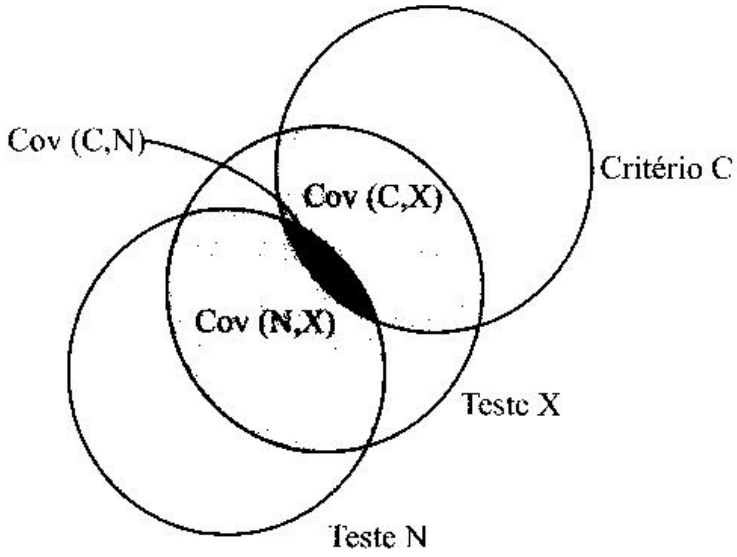


Figura 6-3. Covariância entre critério e dois testes

O uso da *intervenção experimental* aparece como logicamente uma das melhores técnicas para se decidir a validade de construto de um teste. Esta técnica consiste em verificar se o teste discrimina claramente grupos-critério "produzidos" experimentalmente em termos do traço objeto de medida do teste. Assim, um teste que mede ansiedade teria validade de construto (ansiedade) se discriminasse grupo não ansioso de grupo ansioso, definidos estes grupos em termos de manipulações experimentais: o ansioso, por exemplo, criado assim através de experiências provocadoras de ansiedade. Na medida em que se puder garantir que as manipulações feitas nos grupos-critério atingirem exclusivamente o traço em questão, a testagem da hipótese é válida. Como, normalmente, estas manipulações supostamente de uma variável de fato pode afetar uma série de outras variáveis, sobretudo se as variáveis interagirem, fica confusa a decisão sobre em que especificamente os grupos-critério diferem e, consequentemente, fica inconclusiva a decisão sobre a hipótese de que o teste discrimina os grupos-critério exclusivamente em termos do traço que ele pretende medir. Podendo-se garantir que não ocorre tal alastramento das manipulações, a hipótese fica corretamente colocada.

Em conclusão, a técnica da validação de construto via hipótese, que, de um ponto de vista da metodologia científica, se apresenta como a mais direta e óbvia, esbarra na dificuldade que existe na definição inequívoca do critério a ser utilizado como representante da manifestação do traço.

2.4 – A curva de informação da TRI

A TRI trabalha a validade dos testes através de poderosos métodos chamados as funções de informação e de eficiência.

2.4.1 – Função de informação do teste

A informação fornecida pelo teste é simplesmente a soma das informações fornecidas por cada item do mesmo[6], ou seja,

$$I(\theta) = \sum_{j=1}^{n} I_i(\theta) \qquad (6.8)$$

onde, $I_i(\theta)$ é a informação do item, que no capítulo 5 vimos ser ela expressa como $I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)}$. A figura 6-4 ilustra como a função de informação dos itens constitui a função de informação de todo o teste.

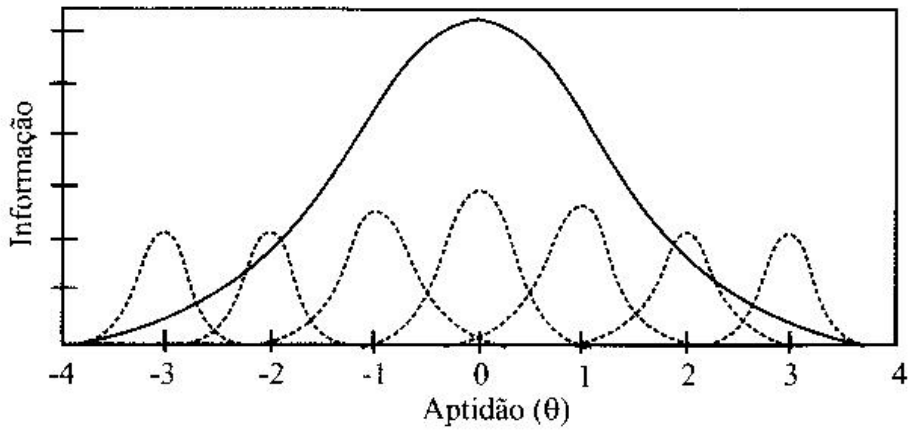


Figura 6-4. Funções de informação de 7 itens (curvas pontilhadas) e do teste

A curva de informação do teste mostra para que faixa de níveis de teta o teste é particularmente válido e para que faixas ele não o é. Veja na

---

6. A função de informação do item foi discutida no capítulo 5.

figura 6-5 que o máximo de informação do teste (teste não verbal de raciocínio para crianças, com 60 itens) é de 11,22 na escala expressa na ordenada e que o valor médio de informação é 7,48.

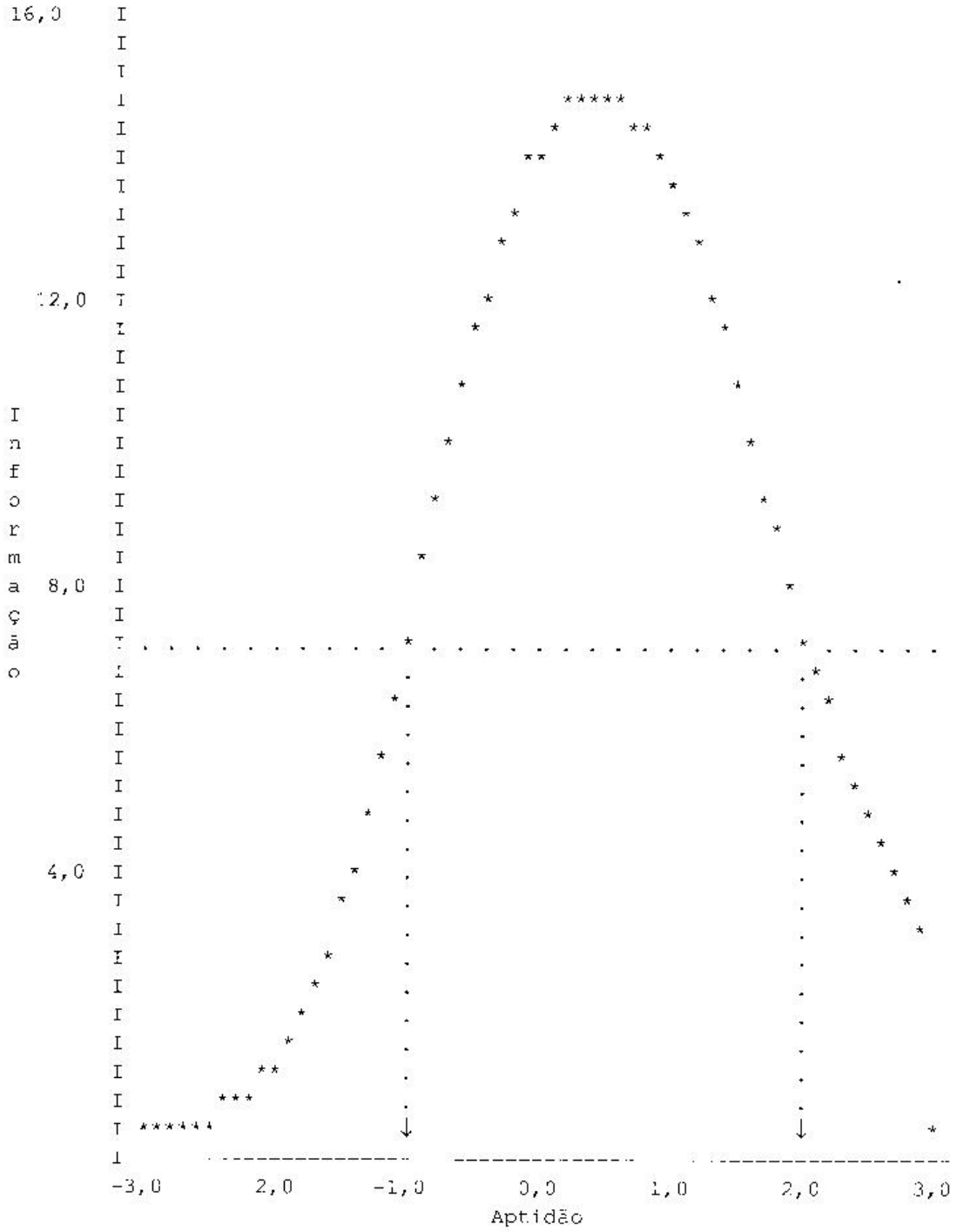


Figura 6-5. Curva de informação de um teste de raciocínio infantil

Tomando este valor de 7,48 como a informação que o teste é capaz de produzir sobre o raciocínio infantil, verificamos que o teste é muito (é particularmente) válido para crianças com aptidão entre -1,0 e +2,0 sigmas (pontos na abcissa marcados com setas), isto é, para crianças que, em raciocínio, se situam entre o percentil 26 e percentil 97, sendo menos válido para crianças abaixo do percentil 26 ou acima do percentil 97. Então, você vê que a curva de informação do teste especifica para que faixa de teta ele é especialmente válido. Ela também mostra que o teste é maximamente válido para crianças com aptidão de raciocínio em torno de 0,5 sigmas, isto é, crianças que se situam no percentil 69 desta habilidade.

Outra maneira de representar esta função de informação do teste é através do erro padrão de medida, chamado na TRI de erro padrão de estimação. A $I(\theta)$, na verdade, é o inverso deste erro:

$$EPE(\theta) = \frac{1}{\sqrt{I(\theta)}} \qquad (6.9)$$

onde, $EPE(\theta)$ = erro padrão de estimação.

Similarmente ao erro padrão de medida da teoria psicométrica clássica, o EPE permite estabelecer intervalos de confiança em torno dos escores θ dos sujeitos.

Exemplo: Um teste de 50 itens aplicado a 100 sujeitos deu $I(\theta) = 4$ para

$\theta = 3$. Para nível de confiança de 95%, qual o intervalo em que cai o θ?

Resposta:

1) para 95%, $z = +/- 1,96$

2) erro máximo = $z_{EPE}(\theta) = z\frac{1}{\sqrt{I(\theta)}} = 1,96\frac{1}{\sqrt{4}} = 0,98$

3) assim, o intervalo para θ será 3 +/– 0,98, isto é, 2,02 a 3,98.

2.4.2 – Ponderação dos itens

Para maximizar, ou seja, dar maior importância a itens que trazem maior informação para a estimação da habilidade dos sujeitos, Birnbaum (1968) e Lord (1980) aconselham dar pesos diferentes a esses itens, atri-

buindo-lhes uma ponderação w_i. Neste caso, a função de informação do teste passaria a ser a seguinte:

$$I_i(\theta) = \frac{[\sum_{i=1}^{n} w_i P_i'(\theta)]^2}{\sum_{i=1}^{n} w_i P_i(\theta) Q_i(\theta)} \qquad (6.10)$$

Este w_i é uma constante k escolhida por você para ponderar itens; ela age diferentemente nos diferentes modelos da TRI; veja como ela atua no modelo de

1-parâmetro: $w_i = k$

2-parâmetros: $w_i = ka_i$

3-parâmetros: $w_i = \frac{ka_i P_i(\theta)}{P_i(\theta)(1 - c_i)}$.

Isto quer dizer que os itens, no modelo de 1-parâmetro (onde se avalia apenas o parâmetro de dificuldade b do item), são todos ponderados da mesma forma, enquanto nos outros modelos eles são ponderados levando em conta os outros parâmetros do item (a e c).

2.4.3 – Função da eficiência relativa

A $I(\theta)$ permite comparar a relativa eficiência de um teste com relação a outro em sua capacidade de estimar a aptidão θ. Veja a fórmula:

$$ER(\theta) = \frac{I_1(\theta)}{I_2(\theta)} \qquad (6.11)$$

Onde

$ER(\theta)$ = eficiência relativa

$I_1(\theta)$ e $I_2(\theta)$ = funções de informação do teste 1 e teste 2.

Exemplo: Se $I_1(\theta) = 30$ e $I_2(\theta) = 15$, então $ER(\theta) = 30/15 = 2$. O teste 1 é duas vezes mais eficiente que o teste 2 para estimar θ, isto é, o teste 1 poderia ser reduzido pela metade ou o teste 2 aumentado 100% para ambos terem o mesmo nível de eficiência.

Concluindo: Deve-se, na verdade, concluir que todas estas técnicas de validação apresentam dificuldades, em particular as técnicas da TCT e

da análise fatorial. Nem por isso se justifica o simples abandono das mesmas. Primeiramente, porque em ciência empírica nada existe de perfeito e isento de erro e, em segundo lugar, a consciência destas dificuldades deve servir para melhorar e não abandonar as técnicas. Aliás, é recomendável o uso de mais de uma das técnicas acima analisadas para demonstrar a validade de construto do teste, dado que a convergência de resultados das várias técnicas constitui garantia para a validade do instrumento.

3 – Validade de critério

Concebe-se como validade de critério de um teste o grau de eficácia que ele tem em predizer um desempenho específico de um sujeito. O desempenho do sujeito torna-se, assim, o critério contra o qual a medida obtida pelo teste é avaliada. Evidentemente, o desempenho do sujeito deve ser medido/avaliado através de técnicas que são independentes do próprio teste que se quer validar.

Costuma-se distinguir dois tipos de validade de critério: (1) validade preditiva e (2) validade concorrente. A diferença fundamental entre os dois tipos é basicamente uma questão do tempo que ocorre entre a coleta da informação pelo teste a ser validado e a coleta da informação sobre o critério. Se estas coletas forem (mais ou menos) simultâneas, a validação será do tipo *concorrente*; caso os dados sobre o critério sejam coletados após a coleta da informação sobre o teste, fala-se em *validade preditiva*. O fato de a informação ser obtida simultaneamente ou posteriormente à do próprio teste não é um fator tecnicamente relevante à validade do teste. Relevante, sim, é a determinação de um critério válido. Aqui se situa precisamente a natureza central deste tipo de validação dos testes, a saber: (1) definir um critério adequado e (2) medir, válida e independentemente do próprio teste, este critério.

Quanto à adequação dos critérios, pode-se afirmar que há uma série destes que são normalmente utilizados, quais sejam:

1) *Desempenho acadêmico*. Talvez seja ou foi o critério mais utilizado na validação de testes de inteligência. Consiste na obtenção do nível de desempenho escolar dos alunos, seja através das notas dadas pelos professores, seja pela média acadêmica geral do aluno, seja pelas honrarias acadêmicas que o aluno recebeu, ou seja, mesmo, pela avaliação puramente subjetiva dos alunos em termos de "inteligente" por parte dos pro-

fessores ou colegas. Embora seja amplamente utilizado, este critério tem igualmente sido muito criticado, não em si mesmo, mas pela deficiência que ocorre na sua avaliação. É sobejamente sabida a tendenciosidade por parte dos professores em atribuir as notas aos alunos, tendenciosidade nem sempre consciente, mas decorrente de suas atitudes e simpatias em relação a este ou aquele aluno. Esta dificuldade poderia ser sanada até com certa facilidade, se os professores tivessem o costume de aplicar testes de rendimento que possuíssem validade de conteúdo, por exemplo. Como esta tarefa é dispendiosa, o professor tipicamente não se dá ao trabalho de validar (validade de conteúdo) suas provas acadêmicas.

Neste contexto, é também utilizado como critério de desempenho acadêmico o *nível escolar* do sujeito: sujeitos mais avançados, repetentes e evadidos. A suposição sendo de que quem continua regularmente ou está avançado academicamente em relação à sua idade possui mais habilidade. Evidentemente, nesta história não entra somente a questão da habilidade, mas muitos outros fatores sociais, de personalidade, etc., tornando este critério bastante ambíguo e espúrio.

2) *Desempenho em treinamento especializado*. Trata-se do desempenho obtido em cursos de treinamento em situações específicas, como no caso de músicos, pilotos, atividades mecânicas ou eletrônicas especializadas, etc. No final deste treinamento há tipicamente uma avaliação, a qual produz dados úteis para servirem de critério de desempenho do aluno. As observações críticas feitas ao ponto 1) valem também neste parágrafo.

3) *Desempenho profissional*. Trata-se, neste caso, de comparar os resultados do teste com o sucesso/fracasso ou o nível de qualidade do sucesso dos sujeitos na própria situação de trabalho. Assim, um teste de habilidade mecânica pode ser testado contra a qualidade de desempenho mecânico dos sujeitos na oficina de trabalho. Evidentemente continua a dificuldade de levantar adequadamente a qualidade deste desempenho dos sujeitos em serviço.

4) *Diagnóstico psiquiátrico*. Muito utilizado para validar testes de personalidade/psiquiátricos. Os grupos-critério são aqui formados em termos da avaliação psiquiátrica que estabelece grupos clínicos: normais vs. neuróticos, psicopatas vs. depressivos, etc. Novamente, a dificuldade continua sendo a adequação das avaliações psiquiátricas feitas pelos psiquiatras.

5) *Diagnóstico subjetivo*. Avaliações feitas por colegas e amigos podem servir de base para estabelecer grupos-critério. É utilizada esta técnica sobretudo em testes de personalidade, onde é difícil encontrar avaliações mais objetivas. Assim, os sujeitos avaliam seus colegas em categorias ou dão escores em traços de personalidade (agressividade, cooperação, etc.), baseados na convivência que eles têm com os colegas. Nem precisa mencionar as dificuldades enormes que tais avaliações apresentam em termos de objetividade; contudo, a utilização de um grande número de juízes poderá diminuir os vieses subjetivos nestas avaliações.

6) *Outros testes* disponíveis. Os resultados obtidos através de outro teste válido, que prediga o mesmo desempenho que o teste a ser validado, servem de critério para determinar a validade do novo teste. Aqui fica a pergunta óbvia: para que criar outro teste se já existe um que mede validamente o que se quer medir? A resposta se baseia numa questão de economia, isto é, utilizar um teste que demanda muito tempo para ser respondido ou apurado como critério para validar um teste que gaste menos tempo.

No caso deste tipo de validade, é preciso atender a duas situações bastante distintas. Primeiramente, quando existem testes comprovadamente validados para a medida de algum traço, eles certamente constituem um critério contra o qual se pode com segurança validar um novo teste. Infelizmente esta situação ocorre quase exclusivamente no caso da medida da inteligência, onde dispomos de alguns testes cuja validade já tem sido comprovada repetidas vezes, como é o caso das escalas de Wechsler (1975), de Stanford-Binet (Terman & Merrill, 1960) e quiçá os dois fatores de inteligência fluida e cristalizada de Cattell (1971) e o fator G de Spearman (1927). Deve-se atender, contudo, que estes testes são válidos somente no original, porque as versões em português ainda não demonstraram suficientemente tal validade. Nos outros campos reina muita confusão. Talvez em personalidade já existam alguns instrumentos válidos, como, por exemplo, o Questionário de Personalidade de Eysenck (*Eysenck Personality Questionnaire* – EPQ, Eysenck & Eysenck, 1975), no que ele se refere às variáveis extroversão e neuroticismo ou ansiedade. O que vale aqui é o princípio de que se houver um teste comprovadamente válido para a medida de algum traço latente, ele certamente pode servir de critério para a validação de um novo teste. Espera-se neste caso que a correlação do novo teste seja elevada, de pelo menos 0,75.

Entretanto, quando não existem testes aceitos como definitivamente validados para avaliar algum traço latente, a utilização desta validação concorrente é extremamente precária. Esta situação infelizmente é a mais comum. De fato, nós temos testes para medir praticamente não importa o quê, como atestam os Buro's *Mental Measurement Yearbooks*, que são publicados periodicamente com centenas e milhares de testes psicológicos existentes no mercado. Neste caso, pode-se utilizar estes testes como critérios de validação, mas o risco é demasiadamente grande, porque se está utilizando como critério testes cuja validade é pelo menos duvidosa.

Pode-se concluir que a validade concorrente só faz sentido se existirem testes comprovadamente válidos que possam servir de critério contra o qual se quer validar um novo teste e que este novo teste tenha algumas vantagens sobre o antigo (como, por exemplo, economia de tempo etc.).

Conclusão geral: Uma pergunta frustrante fica ao final desta exposição sobre validade de critério. Se o pesquisador empregou toda a sua habilidade para construir um teste sob as condições de maior controle possível, por que iria ele validar esta tarefa-teste contra medidas inferiores, representadas pela medida dos vários critérios aqui apresentados. Justifica-se validar medidas supostamente superiores por medidas inferiores, pergunta Ebel (1961)?

Com as críticas de Thurstone (1952) e sobretudo de Cronbach e Meehl (1955), a validade de critério deixou de ser a técnica panacéia de validação dos testes psicológicos em favor da validade de construto. Contudo, estes critérios podem ser considerados bons e úteis para fins de validação de critério. A grande dificuldade em quase todos eles se situa na demonstração da adequação da medida deles; isto é, em geral, a medida dos mesmos é precária, deixando, por isso, muita dúvida quanto ao processo de validação do teste. Entretanto, há exemplos famosos de testes validados através deste método, como é o caso do MMPI.

4 – Validade de conteúdo[7]

Um teste tem validade de conteúdo se ele constitui uma amostra representativa de um universo finito de comportamentos (domínio). É aplicável

---

7. Este tema é amplamente tratado no livro organizado por este autor, *Instrumentos Psicológicos: Manual Prático de Elaboração*, Cap. 7. Brasília, DF: LabPAM/IBAPP, 1999.

quando se pode delimitar *a priori* e com clareza um universo de comportamentos, como é o caso em testes de desempenho, que pretendem cobrir um conteúdo delimitado por um curso programático específico.

Para viabilizar um teste com validade de conteúdo, é preciso que se façam as especificações do teste antes da construção dos itens. 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.

Quanto ao conteúdo, trata-se de detalhá-lo em termos de tópicos (unidades) e subtópicos e de explicitar a importância relativa de cada tópico dentro do teste. Tais procedimentos evitam a super-representação indevida de alguns tópicos e sub-representação de outros por vieses e pendores pessoais do avaliador. Claro que será sempre o avaliador ou equipe de avaliadores que vai definir este conteúdo e a relativa importância de suas partes, mas esta definição deve ser tomada antes da construção dos itens, garantindo certa objetividade, pelo menos, nas decisões.

Quanto aos objetivos, um teste não deve ser elaborado para avaliar exclusivamente um processo. Como na aprendizagem entram em ação vários processos psicológicos, há interesse que todos, ou aqueles que se quer que sejam avaliados por um teste de conteúdo, sejam representados no teste. Por exemplo, o teste deverá conter itens que avaliam a memória (reproduzir), a compreensão (conceituar, definir), a capacidade de comparação (relacionar) e de aplicação dos princípios aprendidos (solucionar problemas, transferência da aprendizagem).

A validade de conteúdo de um teste é praticamente garantida pela técnica de construção do mesmo. Assim, é importante esboçar esta técnica. Ela comporta os seguintes passos:

1 – Definição do domínio cognitivo:

Definir os objetivos ou os processos psicológicos que se quer avaliar. Para esta tarefa é útil se inspirar em alguma taxonomia clássica de objetivos educacionais, como, por exemplo, a taxonomia de Bloom (1956) ou outra. Com base em tal taxonomia, definir os objetivos gerais e específicos que se deseja medir no teste, tais como

– conhecer tais e tais tópicos

– compreender tais e tais tópicos

– aplicar tais e tais tópicos

– analisar tais e tais tópicos.

2 – Definição do universo de conteúdo

Como o teste vai constituir uma amostra representativa do conteúdo, é preciso definir e delimitar o universo do conteúdo programático em termos de divisões e subdivisões (tópicos e subtópicos) ou quantas outras subclassificações forem necessárias. Isto implica em delimitar o conteúdo em suas unidades e subunidades de ensino.

3 – Definição da representatividade de conteúdo

Definir a proporção com que cada tópico e subtópico devem ser representados no teste, decidindo, assim, a importância com que cada um deles aparece no conteúdo total do universo.

4 – Elaboração da tabela de especificação, na qual serão relacionados os conteúdos com os processos cognitivos a avaliar, bem como a importância relativa a ser dada a cada unidade, conforme tabela 6-4.

5 – Construção do teste

Elaborar os itens que irão representar o teste, seguindo as técnicas de construção de itens (Mager, 1981; Pasquali, 1995, 1999).

6 – Análise teórica dos itens

Esta análise visa verificar a compreensão das tarefas propostas no teste por parte dos testandos (análise semântica) e a avaliação da pertinência do item a tal ou tal unidade e avaliando tal ou tal processo cognitivo (análise de juízes).

7 – Análise empírica dos itens

Após a aplicação do teste, os dados podem ser utilizados para uma validação empírica do mesmo para uso futuro. Esta análise implica basicamente na determinação dos níveis de dificuldade e

de discriminação dos itens. A técnica da teoria da resposta ao item (TRI) pode ser de grande valia nesta etapa.

Para facilitar a especificação do teste, pode-se utilizar uma tabela de dupla entrada, com o detalhamento dos objetivos (processos) à esquerda e o detalhamento dos tópicos no topo, explicitando, no corpo da tabela, o número de itens, conforme tabela 6-4.

Tabela 6-4. Tabela de especificação de teste de desempenho (número de itens no corpo)

	Tópicos	1		2			3		
	Subtópicos	1	2	1	2	3	1	2	
	Proporção (%)	10	10	20	5	10	30	25	n
Processos	conhecer	2	2	3	1	2	3	2	15
	compreender	-	1	1	-	1	1	1	5
	aplicar	1	1	1	-	-	2	1	6
	analisar	1	-	1	1	1	-	1	5
n		4	4	6	2	4	6	5	31

A tabela 6-4 explicita que 3 tópicos cobrem o conteúdo total do programa a ser medido (conteúdo programático), tendo cada qual dois ou três subtópicos, cada um com diferentes níveis de representatividade (proporções). Os *n* representam o número de itens por tópico e por processo cognitivo avaliado, estando no corpo da tabela o número de itens que representam cada combinação tópico e processo.

CAPÍTULO 7
Fidedignidade dos testes

I – A teoria

O parâmetro da fidedignidade dos testes vem referenciado sob uma série elevada e heterogênea de nomes. Alguns destes nomes resultam do próprio conceito deste parâmetro, isto é, eles procuram expressar o que ele de fato representa para o teste. Estes nomes são, principalmente, precisão, fidedignidade e confiabilidade. Outros nomes deste parâmetro resultam mais diretamente do tipo de técnica utilizada na coleta empírica da informação ou da técnica estatística utilizada para a análise dos dados empíricos coletados. Entre estes nomes, podemos relacionar os seguintes: estabilidade, constância, equivalência, consistência interna. Esta nomenclatura ficará mais esclarecida quando tratarmos das técnicas estatísticas na determinação do coeficiente de fidedignidade. Em ciências físicas, este parâmetro da medida vem referenciado sob o nome de calibração dos instrumentos.

A fidedignidade ou a precisão de um teste diz respeito à característica que ele deve possuir, a saber, a de medir sem erros, donde os nomes precisão, confiabilidade ou fidedignidade. Medir sem erros significa que o mesmo teste, medindo os mesmos sujeitos em ocasiões diferentes, ou testes equivalentes, medindo os mesmos sujeitos na mesma ocasião, produzem resultados idênticos, isto é, a correlação entre estas duas medidas deve ser de 1. Entretanto, como o erro está sempre presente em qualquer medida, esta correlação se afasta tanto do 1 quanto maior for o erro cometido na medida. A análise da precisão de um instrumento psicológico quer mostrar precisamente o quanto ele se afasta do ideal da correlação 1, determinando um coeficiente que, quanto mais próximo de 1, menos erro o teste comete ao ser utilizado.

O problema da fidedignidade dos testes era tema preferido da Psicometria Clássica, onde a parafernália estatística de estimação deste pa-

râmetro mais se desenvolveu, mas ele perdeu muito em importância dentro da psicometria moderna em favor do parâmetro de validade. De qualquer forma, dentro da TCT o *coeficiente de fidedignidade*, r_{tt}, é definido estatisticamente como a correlação entre os escores dos mesmos sujeitos em duas formas paralelas de um teste, T_1 e T_2. Assim o coeficiente de fidedignidade se define como função da covariância [$Cov(T_1,T_2)$] entre as formas do teste pelas variâncias ($s^2_{T_1}$ e $s^2_{T_2}$) das mesmas, isto é,

$$r_{tt} = \frac{s^2_V}{s^2_T} \qquad (7.1)$$

onde,

r_{tt} : Coeficiente de fidedignidade

s^2_V : Variância verdadeira do teste

s^2_T : Variância total do teste

Esta fórmula surge da definição de precisão ser a função da covariância pela variância, isto é,

$$r_{tt} = \frac{cov}{var} = \frac{\sum t_1 t_2}{N s_1 s_2}$$

Entretanto, em testes paralelos, toda a covariância entre os dois testes é devida unicamente à variância verdadeira, a saber, $\sum v_1 v_2$, que é igual a $\sum v^2$, uma vez que $v_1 = v_2$. Além disso, $s_1 = s_2$ e, consequentemente, o produto das duas variâncias é a variância total dos testes, ou seja, $s_1 s_2 = s^2_T$. Assim, a fórmula se simplifica para

$$r_{tt} = \frac{\sum v^2}{N s^2_T} = \frac{s^2_V}{s^2_T}$$, uma vez que $\frac{\sum v^2}{N} = s^2_V$

O coeficiente de precisão é função da variância verdadeira pela variância total.

Ademais, sendo

$s_T^2 = s_V^2 + s_E^2$ segue que

$s_V^2 = s_T^2 - s_E^2$

Disto resulta que a fórmula 7.1 se torna $r_{tt} = \frac{s_T^2 - s_E^2}{s_T^2}$ ou $r_{tt} = \frac{s_T^2}{s_T^2} - \frac{s_E^2}{s_T^2}$, que, simplificando, dá

$$r_{tt} = 1 - \frac{s_E^2}{s_T^2} \qquad (7.2)$$

onde,

r_{tt} : Coeficiente de fidedignidade

s_E^2: Variância erro da medida

s_T^2: Variância total do teste

Normalmente esta fórmula vem expressa em termos de escores verdadeiros (V) e escores empíricos (T), onde

$$r_{TV} = \sqrt{r_{tt}} = \frac{s_V}{s_T} \qquad (7.3)$$

Esta fórmula surge diretamente da correlação entre o escore empírico total (T) e o escore verdadeiro (V), ou seja,

$r_{TV} = \frac{\sum tv}{N s_T s_V}$ mas, como t = v + e, então

$$r_{TV} = \frac{\sum (v+e)v}{N s_T s_V} = \frac{\sum v^2 + \sum ve}{N s_T s_V}$$

Entretanto, $\frac{\Sigma v^2}{N}$ é a variância de V (s_V^2) e $\frac{\Sigma ve}{N}$ é a covariância entre V e E, a qual, pelo modelo da TCT, é igual a zero.

Disto segue que

$r_{TV} = \frac{s_V^2}{s_T s_V} = \frac{s_V}{s_T}$, que é $\sqrt{r_{tt}}$ pela fórmula 7.1.

Destas fórmulas pode-se observar que, se não houvesse erro na medida, isto é, $s_E^2 = 0$, toda a variância do teste, s_T^2, seria variância verdadeira, s_V^2. Pois, neste caso,

$r_{tt} = \frac{s_V^2}{s_V^2} = 1$ ou $r_{tt} = 1 - \frac{0}{s_T^2} = 1$.

Destas definições também segue o conceito de *erro padrão de medida* (EPM = $\sqrt{s_E^2} = s_E$) ou

$$EPM = s_T\sqrt{1 - r_{tt}} \quad . \qquad \text{(veja capítulo 6)} \qquad (7.4)$$

De fato, de 7.2 sabemos que $r_{tt} = 1 - \frac{s_E^2}{s_T^2}$

Então, $s_E^2 = s_T^2(1 - r_{tt})$, que tirando a raiz quadrada dá a fórmula 7.4.

O problema prático com estas fórmulas é que nem s_V^2 nem s_E^2 são dados pelas respostas dos sujeitos. Assim, o coeficiente de fidedignidade deve ser estimado a partir dos escores empíricos, s_T^2, que é o único dado fornecido por estas respostas; o que será exposto no ponto II.

II – Técnicas de estimação do coeficiente de fidedignidade

Para a estimação da fidedignidade dos testes, existem três tipos de delineamentos (procedimentos experimentais de coleta da informação) e dois tipos ou modelos de análises estatísticas dos dados coletados (correlação e técnicas alfa).

1 – Os delineamentos

1.1 – Uma amostra de sujeitos, um mesmo teste e uma única ocasião

Aplica-se a uma amostra aleatória de sujeitos um teste numa única ocasião e se analisam os dados em termos da consistência interna dos itens através de análises estatísticas, a saber: (1) análise das duas metades ou (2) análise pelas técnicas alfa, detalhadas mais adiante.

1.2 – Uma amostra de sujeitos, dois testes e uma única ocasião

Aplica-se a uma amostra aleatória de sujeitos duas formas paralelas do teste ou dois testes paralelos (T1 e T2) numa única ocasião e faz-se a análise da correlação entre as distribuições dos dois testes ou formas paralelas.

1.3 – Uma amostra de sujeitos, um mesmo teste e duas ocasiões

Aplica-se a uma amostra aleatória de sujeitos um teste na ocasião 1 (O1) e reaplica-se à mesma amostra o mesmo teste em uma ocasião ulterior (O2) e faz-se a correlação entre os dois conjuntos de dados.

2 – Técnicas estatísticas

Embora haja dezenas de índices de fidedignidade (Gulliksen, 1950; Nunnally, 1978), há basicamente duas técnicas estatísticas para a estimação do coeficiente de fidedignidade: a correlação simples e a(s) técnica(s) alfa. A primeira trabalha com a correlação e as segundas com a variância.

A apresentação de todos estes índices seria de pouca utilidade e não iluminaria profundamente os esclarecimentos nesta área estatisticamente complexa da fidedignidade para o pesquisador e profissional. Para quem quiser se enredar nesta malha de coeficientes, aconselhamos o estudo dos dois autores acima referidos. Aqui apresentaremos somente alguns índices que são os mais utilizados na prática da pesquisa e profissional.

2.1 – A correlação

Para obter o coeficiente de fidedignidade através do modelo da correlação, utilizam-se os dados coletados via um dos três procedimentos

ou técnicas empíricas de coleta da informação (delineamentos). Cada um dos delineamentos acima expostos apresenta vantagens e desvantagens. Por serem diferentes os procedimentos empíricos utilizados por estes delineamentos, os coeficientes de fidedignidade que deles resultam tomam nomes diferentes, embora estejam querendo falar da mesma coisa, como se vê na tabela 7-1.

O coeficiente de correlação expressa o nível de relação ou correspondência que existe entre dois eventos, tais como evento *i* e evento *j*, e vem designado como r_{ij}. É o tipo de coeficiente utilizado com as três técnicas experimentais de estimação da fidedignidade acima descritas. Por exemplo, se aplico um teste a 100 sujeitos num dia e o aplico novamente depois de corridos 30 dias, e se a classificação dos 100 sujeitos for idêntica nas duas ocasiões, isto é, os mais fortes saíram os mais fortes e os mais fracos os mais fracos em ambas as aplicações, existe uma correspondência de 100% entre as duas aplicações e o coeficiente de correlação será de 1, isto é, perfeito. Obviamente, nunca acontecerá que o primeiro colocado na primeira aplicação do teste seja o primeiro na segunda, o segundo colocado seja também o segundo na segunda aplicação e assim até o último, porque, pelo menos, os erros que se cometem na medida na primeira e na segunda aplicação não são os mesmos, dado que eles são aleatórios. Mas se a classificação dos sujeitos se mantém mais ou menos a mesma em ambas as aplicações, continua havendo uma correspondência entre estas aplicações. Qual é o montante desta correspondência é o que o coeficiente de correlação determina, ou seja, quanto mais próximo de 1 positivo seja este coeficiente, mais próxima de 100% será a correspondência direta entre as duas classificações; quanto mais próxima de -1 (negativo) a correlação, tanto mais certeza temos que os primeiros classificados na primeira aplicação do teste serão os últimos na segunda e vice-versa. Um coeficiente de 0 indicaria que não haveria nenhuma relação entre o que aconteceu na primeira aplicação e na segunda; isto é, a partir da classificação dos sujeitos obtida na primeira aplicação não dá para predizer nada sobre como será esta classificação na segunda aplicação.

Tabela 7-1. Estimação do coeficiente de fidedignidade segundo os três delineamentos

Técnica	Procedimento empírico	Coeficiente	Vantagem	Desvantagem
Formas paralelas	Aplicar 2 formas paralelas de um teste, T_1 e T_{12}, a amostra representativa de sujeitos e calcular a correlação r_{T1T2}	Equivalência	• bate diretamente com o conceito de fidedignidade • testes aplicados numa só ocasião	Difícil conseguir formas perfeitamente paralelas, i.é, medir mesmo traço latente com itens diferentes
Teste-reteste	Aplicar mesmo teste T aos mesmos sujeitos em duas ocasiões diferentes, O_1 e O_2, e calcular correlação r_{O1O2}	Estabilidade ou constância	Garantia da equivalência (paralelismo), pois se trata do mesmo teste T	– Difícil definir intervalo ideal de tempo entre O_1 e O_2 – Difícil controlar eventos que ocorrem entre O_1 e O_2
Duas metades	Aplicar um teste T a amostra representativa; dividir T em 2 partes equivalentes (itens pares vs. ímpares; 1^a e 2^a metades; ou outra divisão); calcular correlação entre as 2 metades; aplicar a correção de Spearman-Brown	Consistência interna do teste	Exige apenas uma aplicação (evita eventos temporais)	– Difícil garantir equivalência das 2 metades – Técnica de 1^a e 2^a metades não controla fadiga dos testandos ao tomar 2^a metade e normalmente esta metade contém os itens mais difíceis

A tabela 7-2 mostra como se calcula o coeficiente de correlação entre os escores de 10 sujeitos aos quais foram aplicadas duas formas paralelas de um teste de raciocínio verbal (RV1 e RV2 – dados fictícios), onde x e y são os desvios de X e de Y em relação às suas respectivas médias.

Tabela 7-2 – Cálculo do coeficiente de correlação

	RV1	RV2					
Sujeito	X	Y	x	y	x^2	y^2	xy
1	15	18	1	3	1	9	3
2	10	11	-4	-4	16	16	16
3	12	14	-2	-1	4	1	2
4	20	18	6	3	36	9	18
5	16	16	2	1	4	1	2
6	14	15	0	0	0	0	0
7	11	13	-3	-2	9	4	6
8	09	12	-5	-3	25	9	15
9	18	15	4	0	16	0	0
10	15	18	1	3	1	9	3
Soma	140	150			112	58	65
Média	14,00	15,00					

$$s_X = \sqrt{\frac{\sum x^2}{N}} = \sqrt{\frac{112}{10}} = \sqrt{11,2} = 3,35 \qquad s_Y = \sqrt{\frac{58}{10}} = \sqrt{5,8} = 2,41$$

$$r_{XY} = \frac{\sum xy}{N s_X s_Y} = \frac{65}{10 \times 3,35 \times 2,41} = \frac{65}{80,74} = 0,81$$

O coeficiente de correlação de 0,81 é bastante elevado, indicando que 65,61% (= $0,81^2$) da classificação que ocorreu no RV1 foi idêntica ao

que ocorreu no RV2. Quando o coeficiente de correlação representa o coeficiente de fidedignidade, este deve se aproximar de 1 para se poder afirmar que o teste é preciso. Não é suficiente, neste caso, que os coeficientes sejam estatisticamente significativos; eles devem se aproximar de 1. A razão disto se situa no fato de que estamos trabalhando com o conceito de testes paralelos que, por definição, devem ter médias e variâncias iguais. Assim, a utilização de dois testes paralelos para avaliar os mesmos sujeitos deve produzir resultados idênticos, descontados os erros de medida; isto é, a correlação entre os dados dos dois testes deve ser próxima de 1. Em outras palavras, não é bastante que os dois testes estejam relacionados; eles devem ser idênticos em seus resultados. De fato, um coeficiente de fidedignidade abaixo de 0,80 já é fraco e um de 0,70 já é inaceitável. Na verdade, um coeficiente de 0,70 representa uma covariância de apenas 49% (= $0{,}70^2$) entre os resultados dos dois testes e um erro de 51%. Isto não é aceitável se queremos afirmar que os dois testes são paralelos. Consequentemente, coeficientes em torno de 0,90 ou maiores são normalmente os esperados para expressar a fidedignidade de um teste. Coeficientes em torno de 0,80 usualmente são considerados razoáveis, enquanto coeficientes de precisão abaixo de 0,70 não são normalmente suficientes como demonstração de uma fidedignidade aceitável para um teste.

No caso da técnica das duas metades, a correlação entre os dados das metades deve ser corrigida através da fórmula de Spearman-Brown. Esta correção se impõe dado que a correlação se baseia somente na metade do tamanho da escala; o comprimento do teste afeta bastante a coeficiente de precisão (isto será explicitado mais adiante no ponto III). O exemplo da tabela 7-3 ilustra o cálculo da correção de Spearman-Brown, onde m_1 e m_2 são os desvios de M_1 e M_2 em relação às suas respectivas médias.

Tabela 7-3. Cálculo do coeficiente de precisão com a técnica das duas metades

Sujeitos	T	M_1	M_2	m_1	m_2	m_1^2	m_2^2	m_1m_2
1	50	30	20	9,2	0	84,64	0	0
2	40	20	20	-0,8	0	0,64	0	0
3	46	22	24	1,2	4	1,44	16	4,8
4	38	20	18	-0,8	-2	0,64	4	1,6
5	30	14	16	-6,8	-4	46,24	16	27,2
6	42	20	22	-0,8	2	0,64	4	-1,6
7	36	18	18	-2,8	-2	7,84	4	5,6
8	32	17	15	-3,8	-5	14,44	25	19,0
9	44	21	23	0,2	3	0,04	9	0,6
10	50	26	24	5,2	4	27,04	16	20,8
Soma		208	200	0	0	183,60	94	78
Média		20,8	20,0					

$$s_{M_1} = \sqrt{\frac{\sum m_1^2}{N}} = \sqrt{\frac{183,6}{10}} = 4,28 \quad s_{M_2} = \sqrt{\frac{\sum m_2^2}{N}} = \sqrt{\frac{94}{10}} = 3,07$$

$$r_{12} = \frac{\sum m_1 m_2}{N s_1 s_2} = \frac{78}{10 \times 4,28 \times 3,07} = 0,59$$

Correção Spearman-Brown:

$$r_{tt} = \frac{n r_{12}}{1 + (n-1) r_{12}} = \frac{2 \times 0,59}{1 + 0,59} = 0,74$$

De fato, vê-se que a correção de Spearman-Brown elevou o índice de 0,59 para 0,74; coeficiente, obviamente, que ainda deixa a desejar, embora o número de sujeitos em que ele foi baseado seja ridiculamente pequeno.

Crítica. Todas estas técnicas e delineamentos apresentam problemas, que à vezes podem ser difíceis de serem superados. Por exemplo, no caso das formas paralelas, o problema mais grave se situa na dificuldade de se conseguirem formas perfeitamente paralelas. Teoricamente isto não é impossível, ainda que laborioso. Na verdade, se tivermos à disposição uma série grande de itens que sabemos medir um traço qualquer e dos quais conhecemos os principais parâmetros, a saber, os índices de dificuldade e de discriminação, podemos, então, construir duas ou mais formas paralelas pareando os itens em termos destes parâmetros, a saber, o primeiro item da forma *A* tendo os mesmos índices de dificuldade e de discriminação da forma *B* e assim por diante. A teoria da resposta ao item (TRI) pode ser de grande valia neste empenho, dado que ela nos dá estes parâmetros dos itens. Além desse problema, existe, no caso das formas paralelas, a situação em que a segunda forma aplicada pode ser influenciada pela aplicação da primeira. Isto é particularmente grave em situações em que o teste constitui uma instância de aprendizagem, como é o caso quando as tarefas envolvidas no teste se apresentam como novidade para o sujeito. Neste caso, as respostas à segunda forma terão também a influência desta aprendizagem ou desta exposição à novidade. Isto quer dizer que o sujeito ao tomar a forma *B* já não é o mesmo sujeito, pois a exposição à situação *A* o modificou. Isso não seria problema grave se todos os sujeitos fossem afetados do mesmo jeito pela tomada do primeiro teste, o que obviamente é difícil de ocorrer e aceitar.

No caso do teste-reteste, três tipos de dificuldades podem ocorrer: (1) particularmente no caso de testes curtos e no caso do intervalo entre teste e reteste ser curto, a memória pode entrar em jogo, fazendo com que o sujeito dê a mesma resposta que deu antes simplesmente por se lembrar dela e não em função de sua reação/conhecimento atuais; (2) em caso de testes longos, entra uma questão de atitude, isto é, o sujeito pode se chatear, se irritar ou reagir de qualquer outra forma negativa contra ter que repetir a mesma monótona ladainha de responder ao mesmo enorme número de itens que acabara de responder; (3) particularmente no caso do intervalo entre teste e reteste ser longo, uma série de fatores pessoais e do meio ambiente pode mudar e, consequentemente, o sujeito responderá afetado

por estes novos fatores internos ou externos que mudaram com o passar do tempo (cf. Campbell & Stanley, 1973).

Para fugir dos problemas de teste-reteste, utiliza-se a técnica das duas metades para calcular o coeficiente de fidedignidade. Ela consiste em dividir o teste em duas metades equivalentes (ou quantas partes desejadas). Mas aí está o problema, as metades têm que ser equivalentes, isto é, os itens têm que ser em número igual nas duas metades, ter o mesmo nível de dificuldade, ter o mesmo nível de discriminação e ter os mesmos índices de consistência interna. São exigências demais para se poder assumi-las com tranquilidade. Além disso, no caso da divisão em primeira metade e segunda metade, há necessidade de se mostrar que as respostas aos itens da primeira metade não afetaram as da segunda. E ainda existe o problema, em teste onde os itens são ordenados em termos de dificuldade, de que a segunda metade é mais difícil do que a primeira. Isto não será problema maior se em cada metade os itens forem rigorosamente ordenados em termos de dificuldade, isto é, o item mais fácil de primeira metade emparelha com o mais fácil de segunda, o item mais difícil de primeira emparelha com o mais difícil da segunda e assim por diante. Novamente, um caso difícil de se conseguir na prática. Ademais, no caso da divisão do teste em metades, é preciso corrigir o índice estatístico (correlação entre os escores das duas metades) pela fórmula de Spearman-Brown, dado que o comprimento do teste afeta substancialmente o coeficiente de fidedignidade (cf. ponto III mais adiante).

Note que todos estes problemas não tornam estas técnicas inúteis; apenas o pesquisador deve demonstrar que, no seu caso, eles não ocorreram ou que sua ocorrência não teve influência importante.

2.2 – Coeficientes alfa (α)

Há uma série de técnicas de estimativa de coeficientes de precisão que resultam da análise estatística dos dados de uma única aplicação de um teste a uma amostra representativa de sujeitos. Eles visam verificar a consistência interna do teste através da análise da consistência interna dos itens, isto é, verificando a congruência que cada item do teste tem com o restante dos itens do mesmo teste. O caso mais geral deste tipo de análises é o coeficiente alfa de Cronbach, tendo como casos particulares uma série de outros coeficientes, tais como o de Rulon, Guttman-Flanagan e Kuder-Richardson. Consideremos alguns deles a seguir.

2.2.1 – O coeficiente alfa

Embora não tenha sido o primeiro a trabalhar com este tipo de análise, contudo foi Cronbach (1951) quem propôs este coeficiente geral que reflete o grau de covariância dos itens entre si, servindo assim de indicador da consistência interna do próprio teste. Sua fórmula é a seguinte:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum s_i^2}{s_T^2}\right) \qquad (7.5)$$

onde

n: número de itens

$\sum s_i^2$: soma das variâncias dos *n* itens

s_T^2: variância total dos escores do teste.

A obtenção do coeficiente alfa demanda o cálculo de três parâmetros, a saber: a variância total do teste (s_T^2), a variância de cada item individualmente (s_i^2) e a soma das variâncias destes itens ($\sum s_i^2$). Esta fórmula deixa entrever que será maior o índice alfa quando a variância específica de cada item for pequena e for grande a variância que eles em conjunto produzem. Significa que a soma das variâncias dos itens individuais se reduz e aumenta a variância que eles têm em comum, que é aquela que garante a congruência (consistência interna) entre os itens do mesmo teste. Isto significa que, no extremo, toda a variância produzida pelo teste (s_T^2) seria covariância. Assim, a fórmula de Cronbach mostra que, se todos os itens variarem do mesmo jeito, isto é, se não houver variância entre os itens individualmente, o alfa será igual a 1; quer dizer, que os itens serão totalmente homogêneos, de fato idênticos, produzindo exatamente a mesma variância. Como tal evento não é de se esperar, o alfa dará o tanto de congruência ou covariação que os itens têm dentro do teste. O coeficiente alfa vai de 0 a 1, o 0 indicando ausência total de consistência interna dos itens e o 1, presença de consistência de 100%.

Uma pausa. Como é que a consistência interna de um conjunto de itens (isto é, um teste) fala da precisão de medida dos mesmos? A lógica

atrás dessa suposição é de que quanto menos variabilidade um mesmo item produz numa amostra de sujeitos, menos erros ele provoca. Assim, quanto menor a variância do item, mais preciso é o item. Thurstone (1927) já definia a precisão do item pelo fato dele produzir uma posição clara (um processo descriminante modal, dizia ele) numa escala e tendo um desvio padrão reduzido. Por isso, quanto menor a soma das variâncias dos itens (o numerador na fórmula de Cronbach), mais consistente, portanto, mais preciso é o teste.

Para ilustrar como se calcula o alfa, veja o exemplo na tabela 7-4.

Tabela 7-4. Cálculo do coeficiente alfa de Cronbach

	Itens								
Sujeitos	1	2	3	4	5	6	T	t	t^2
1	1	1	0	1	0	0	3	-0,2	0,04
2	1	0	1	1	1	0	4	0,8	0,04
3	0	1	1	0	0	0	2	-1,2	1,44
4	1	1	1	1	1	1	6	2,8	7,84
5	1	0	0	0	0	0	1	-2,2	4,84
Soma	4	3	3	3	2	1	16		14,80
Média	0,8	0,6	0,6	0,6	0,4	0,2	3,2		
Variância	0,16	0,24	0,24	0,24	0,24	0,16			2,96

$$s_T^2 = \frac{\sum t^2}{N} = \frac{14,8}{5} = 2,96$$

$$s_i^2 = 0,16 + 0,24 + 0,24 + 0,24 + 0,24 + 0,16 = 1,28$$

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum s_i^2}{s_T^2}\right) = \frac{6}{6-1}\left(1 - \frac{1,28}{2,96}\right) = 0,681$$

1 = resposta correta; 0 = resposta errada; $t = T - \overline{T}$

Para efetuar os cálculos acima, é preciso computar primeiro a variância de cada item individualmente. Este cálculo está ilustrado na tabela 7-5, na qual é calculada a variância do item 1.

Tabela 7-5. Cálculo da variância do item 1

Sujeitos	Item 1 (X)	x	x^2
1	1	0,2	0,04
2	1	0,2	0,04
3	0	-0,8	0,64
4	1	0,2	0,04
5	1	0,2	0,04
Soma	4	0,0	0,80
Média	0,8		

$$s_I^2 = \frac{\sum x^2}{N} = \frac{0,80}{5} = 0,16 \text{ ou}$$

$$s_I^2 = pq = \frac{4}{5} x \frac{1}{5} = 0,16$$

2.2.2 – Casos particulares de alfa

Historicamente estes casos especiais de alfa surgiram antes do próprio alfa de Cronbach (1951) na Psicometria para estimar a consistência interna de um teste. É o caso de Kuder-Richardson (1937), Flanagan (1937), Rulon (1939) e Guttman (1945), entre outros. Eles são, contudo, apresentados após, dado que logicamente constituem apenas casos específicos da fórmula geral de alfa de Cronbach. Vamos ver brevemente alguns destes casos, para os quais apresentaremos, logo em seguida à sua exposição, um exemplo comum para todos eles, nas tabelas 7-5 a 7-7.

1) – Rulon:

$$r_{tt} = 1 - \frac{s_D^2}{s_T^2} \qquad (7.6)$$

onde,

s_D^2 é a variância das diferenças entre os escores dos sujeitos nas duas metades de um teste.

s_T^2 é a variância total dos sujeitos no teste.

As diferenças *D* entre os escores das duas metades, metades que se supõem paralelas, seriam devidas somente ao erro de medida. Se não houvesse erro, *D* seria 0 e a correlação seria 1 – 0 = 1. A fórmula de Rulon surge diretamente da fórmula 7.2, isto é, $r_{tt} = 1 - \frac{s_E}{s_T^2}$, onde a variância erro (s_E^2) é substituída pela variância das diferenças (s_D^2).

2) – Guttman-Flanagan:

$$r_{tt} = 2\left(1 - \frac{s_P^2 + s_I^2}{s_T^2}\right) \qquad (7.7)$$

onde,

s_P^2: variância dos escores nos itens pares

s_I^2: variância dos escores nos itens ímpares

s_T^2 : variância total do teste

Tanto o caso Rulon quanto o de Guttman-Flanagan constituem casos particulares de alfa quando n = 2, isto é, as duas metades de um teste, onde s_1^2 e s_2^2 são substituídos por s_D^2 em Rulon e por s_P^2 e s_I^2 em Guttman-Flanagan. Neste caso, $\alpha = \frac{2}{2-1}\left(1 - \frac{s_I^2 + s_2^2}{s_T^2}\right)$.

O problema em dividir o teste em duas metades consiste em saber em que duas metades ele deve ser dividido. A resposta correta seria a seguinte: duas metades que sejam equivalentes. O alfa de Cronbach é outra solução para o problema, visto que ele representa o valor médio de todas as metades possíveis em que um teste de *n* itens pode ser dividido, representando, portanto, o valor esperado das metades, a saber, $\alpha = E\left(\frac{\alpha}{2}\right)$. Note-se, contudo, que um teste de apenas 10 itens, por exemplo, terá 252 metades possíveis!

3) – Kuder-Richardson

São as famosas fórmulas de número 20 e 21 destes dois autores.

$$KR_{20} = \frac{n}{n-1}\left(1 - \frac{\sum pq}{s_T^2}\right) \qquad (7.8)$$

Esta é a fórmula de alfa quando os itens são dicotômicos, caso no qual a variância é $s_i^2 = pq$, onde *p* representa a proporção de sujeitos que acertam o item e *q* a proporção dos que o erram.

A outra fórmula é:

$$KR_{21} = \frac{n}{n-1}\left(1 - \frac{\overline{T} - \left(\frac{\overline{T}^2}{n}\right)}{s_T^2}\right)$$

Esta fórmula supõe que, além de dicotômicos, os itens têm o mesmo nível de dificuldade, onde

$$\sum pq = npq = np(1-p) = np - npp = np - \frac{nppn}{n} = \overline{T} - \frac{\overline{T}^2}{n}.$$

Tabela 7-6. Exemplo geral para estas várias fórmulas, onde 1 significa que o item foi acertado e 0 foi errado (Muñiz, 1992: 53)

Sujeitos	Itens						T	P	I	P-I (D)	T^2	P^2	I^2	D^2
	1	2	3	4	5	6								
1	1	1	0	1	0	0	3	2	1	1	9	4	1	1
2	1	0	1	1	1	0	4	1	3	-2	16	1	9	4
3	0	1	1	0	0	0	2	1	1	0	4	1	1	0
4	1	1	1	1	1	1	6	3	3	0	36	9	9	0
5	1	0	0	0	0	0	1	0	1	-1	1	0	1	1
Soma	4	3	3	3	2	1	16	7	9	-2	66	15	21	6
Média							3,2	1,4	1,8	-0,4				

T = soma dos itens acertados pelo sujeito (= escore total); P = itens pares acertados; I = itens ímpares acertados;

D = diferença entre acertos nos itens pares e nos itens ímpares

Com os dados da tabela 7-6 podemos calcular as médias e as variâncias que vamos precisar para o cômputo dos coeficientes de fidedignidade utilizando as várias fórmulas acima especificadas. Os cálculos das médias e variâncias estão na tabela 7-7 e dos coeficientes de precisão estão na tabela 7-8.

Tabela 7-7. Cálculo das médias e das variâncias dos dados da tabela 7-6

Média Total	$\overline{T} = \frac{\sum T}{N} = \frac{16}{5} = 3,2$		
Variância	Fórmula	Dados	Resultados
Total	$s_T^2 = \frac{\sum T^2}{N} - \overline{T}^2$	$\frac{66}{5} - 3,2^2$	2,96
Dos Itens			
item 1	$s_i^2 = p_i q_i$	4/5 x 1/5	0,16
item 2		3/5 x 2/5	0,24
item 3		3/5 x 2/5	0,24
item 4		3/5 x 2/5	0,24
item 5		2/5 x 3/5	0,24
item 6		1/5 x 4/5	0,16
Itens pares	$s_P^2 = \frac{\sum P^2}{N} - \overline{P}^2$	$15/5 - 1,4^2$	1,04
Itens impares	$s_I^2 = \frac{\sum I^2}{N} - \overline{I}^2$	$21/5 - 1,8^2$	0,96
Diferença	$s_D^2 = \frac{\sum D^2}{N} - \overline{D}^2$	$6/5 - (-0,4)^2$	1,039

p_i = proporção de acertos do item; q_i = complementar de p_i (= 1 – p_i).

O cálculo dos coeficientes de precisão dos dados da tabela 7-6, utilizando as várias fórmulas descritas acima, se encontra na tabela 7-8.

Tabela 7-8. Cálculo dos coeficientes de precisão dos dados da Tabela 7-6

Coeficiente	Fórmula	Dados	Resultados
Alfa (α)	$\frac{n}{n-1}\left(1-\frac{\sum s_i^2}{s_T^2}\right)$	$\frac{6}{6-1}\left(1-\frac{,16+,24+,24+,24+,24+,16}{2,96}\right)$	0,681
Rulon	$1-\frac{s_D^2}{s_T^2}$	$1-\frac{1,039}{2,96}$	0,648
Guttman-Flanagan	$2\left(1-\frac{s_P^2+s_I^2}{s_T^2}\right)$	$2\left(1-\frac{1,04+0,96}{2,96}\right)$	0,648
KR_{20}	$\frac{n}{n-1}\left(1-\frac{\sum pq}{s_T^2}\right)$	$\frac{6}{6-1}\left(1-\frac{,16+,24+,24+,24+,24+,16}{2,96}\right)$	0,681
KR_{21}	$\frac{n}{n-1}\left(1-\frac{\overline{T}-\frac{\overline{T}^2}{n}}{s_T^2}\right)$	$\frac{6}{6-1}\left(1-\frac{3,2-\frac{3,2^2}{6}}{2,96}\right)$	0,594

Observa-se que o coeficiente de precisão é praticamente o mesmo, qualquer que seja a fórmula estatística utilizada para o seu cálculo. Apenas não é isto verdade para o caso da fórmula KR_{21}, que produz um coeficiente bem menor que o KR_{20}, e isto porque os itens não têm o mesmo nível de dificuldade, condição necessária para o uso da fórmula KR_{21}. Portanto, o uso do KR_{21} no caso não seria justificável.

À guisa de informação, note que há outras maneiras de calcular o alfa, salientado-se entre elas, sobretudo:

a) o cálculo de alfa mediante a análise da variância (Hoyt, 1941; Winer, 1971: 283-296);

b) o cálculo de alfa baseado na análise fatorial. Dois índices são aqui os mais utilizados: o coeficiente teta (θ) de Carmines (Carmines & Zeller, 1979) e o Omega (Ω) de Heise e Bohrnstedt (1970).

Enfim, há uma seara imensa de maneiras estatísticas de trabalhar a precisão através da análise da consistência interna dos itens.

2.3 Estimação da fidedignidade de uma bateria de testes

Proposto por Raju (1977), o beta constitui um caso generalizado de alfa, pois é utilizado no caso de uma bateria de testes, da qual se possuem apenas os escores totais de cada subteste (e não os dados de cada item), para a qual se quer obter uma estimação do coeficiente alfa baseado nos dados dos subtestes como se estes fossem itens. O beta será idêntico ao alfa se todos os subtestes tiverem o mesmo número de itens, caso contrário ele representa uma estimação melhor do coeficiente de precisão do que o próprio alfa, pois este subestima este coeficiente em caso de desigualdade de número de itens nos subtestes. A fórmula é a seguinte:

$$\beta = \frac{s_T^2 - \sum s_j^2}{s_T^2 \left[1 - \sum \left(\frac{n_j}{n} \right)^2 \right]} \qquad (7.9)$$

onde,

Os somatórios são feitos sobre o número de subtestes na bateria

s_T^2 é a variância total da bateria

s_j^2 é a variância de cada subteste

n_j é o número de itens em cada subteste

n é o número total dos itens de toda a bateria

Exemplo:

Bateria de 3 testes de raciocínio (verbal, numérico, espacial) com 10, 20 e 40 itens cada (total de itens = 70), variância total = 40 e variâncias para cada subteste de 8, 10 e 12. Qual o alfa e o beta da bateria?

Resposta:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{s_j^2}{s_T^2}\right) = \frac{3}{3-1}\left(1 - \frac{8+10+12}{40}\right) = 0{,}375$$

$$\beta = \frac{40 - (8+10+12)}{40\left[1 - \left\{\left(\frac{10}{70}\right)^2 + \left(\frac{20}{70}\right)^2 + \left(\frac{40}{70}\right)^2\right\}\right]} = 0{,}439$$

Com a correção de Raju, o alfa passa de 0,375 para 0,439, visto que os subtestes possuem número diferente de itens. Se estes tivessem o mesmo número de itens, o beta seria igual a alfa, pois $n = kn_j$.

3 – Casos específicos

3.1 – Estimação do escore verdadeiro

Conhecendo-se o coeficiente de fidedignidade, calculado por alguma das técnicas acima expostas, é possível se fazer alguma estimação de qual seria o escore verdadeiro (V) dos sujeitos no teste e, assim, delimitar o tamanho do erro (E) existente no escore total (T). As técnicas mais conhecidas para esta estimação são as três expostas a seguir.

3.1.1 – Desigualdade de Chebychev

A fórmula geral de Chebychev é a seguinte:

$$DC = \forall K P\left\{\left|X - \overline{X}\right| \leq K(s_X)\right\} \geq 1 - \frac{1}{K^2} \qquad (7.10)$$

a qual, em terminologia da Psicometria Clássica, reza

$$DC = \forall K P\left\{\left|T - V\right| \leq K(s_E)\right\} \geq 1 - \frac{1}{K^2} \qquad (7.11)$$

Esta versão psicométrica da fórmula surge do fato de que a média dos escores empíricos é o escore verdadeiro, isto é, E(T) = V e o desvio padrão de T é o desvio padrão do erro, dado que, fixado o V, toda a variabilidade de T é devida unicamente ao erro, resultando em que $s^2(T|V) = s^2(E|V) = s_E^2$.

O valor K resulta do nível de confiança que se quer ter do intervalo dentro do qual se situará o valor do escore verdadeiro. Este nível é tipicamente de 99% ou 95%. O cálculo do K se faz tornando o termo $1 - \frac{1}{K^2}$ igual ao nível de confiança. Assim, se $1 - \frac{1}{K^2} = 0{,}99$ (nível de confiança de 99%), então

$$\frac{1}{K^2} = 1 - 0{,}99$$

$$K^2 = \frac{1}{1 - 0{,}99}$$

$$K = \sqrt{\frac{1}{1 - 0{,}99}} = 10$$

A leitura desta equação é a seguinte: Para um dado valor de K, a probabilidade do escore V cair num intervalo delimitado é definida pelo erro padrão do erro e pelo nível de confiança escolhido.

Exemplo: 1.000 sujeitos se submeteram ao teste T, cujo coeficiente de fidedignidade do teste = 0,75, e obtiveram média = 50 com desvio padrão = 10. Qual será, a um nível de confiança de 90%, o escore verdadeiro do sujeito que obteve um escore empírico de 100?

Temos, então, os seguintes dados:

N = 1.000 $\quad s_T = 10 \quad r_{tt} = 0{,}75 \quad \overline{T} = 50 \quad T = 100$

Assim, o

$$s_E = s_T\sqrt{1 - r_{tt}} = 10\sqrt{1 - 0{,}75} = 10\sqrt{0{,}25} = 5$$

$$K = \sqrt{\frac{1}{1 - 0{,}90}} = \sqrt{10} = 3{,}16$$

Substituindo estes valores na equação 7.10, temos

$$P\{|100 - V| \leq 3{,}16x5\} \geq 0{,}90$$
$$P\{|V - 100| \leq 3{,}16x5\} \geq 0{,}90$$
$$P\{-15{,}8 \leq V - 100 \leq 15{,}8\} \geq 0{,}90$$
$$P\{84{,}2 \leq V \leq 115{,}8\} \geq 0{,}90$$

Note que 15,8 é Ks_E, isto é, 3,16 x 5; 84,2 é a diferença 100 - (3,16 x 5).

Assim, para sujeitos com escore empírico de 100 no teste T, o seu escore verdadeiro (V) se situa entre 84,2 e 115,8, uma amplitude de 31,6 pontos. Esta faixa enorme se deve a que o coeficiente de fidedignidade do teste (= 0,75) deixa a desejar.

3.1.2 – Estimação do V via distribuição normal dos erros

Faz-se a suposição de que os escores empíricos (T) e os erros de medida (E) se distribuem normalmente para um escore verdadeiro (V), o que implica que se distribuem segundo a curva normal. Sendo isto verdadeiro, segue que

$$f(E|V) \approx N(0|s_E^2) \text{ e dado que } T = V + E$$
$$f(T|V) \approx N(V|s_E^2)$$

Esta técnica de estimação faz duas suposições: (1) a normalidade na distribuição dos escores e, (2) a homoscedasticidade, isto é, que as variâncias são idênticas para todos os escores ao longo da escala. A primeira suposição é normalmente considerada razoável de se fazer, mas a segunda é muito criticada, sobretudo para escores extremos da escala (Feldt, Steffan & Gupta, 1985).

De qualquer forma, se as suposições se mantêm, os dados do exemplo acima (ponto 3.1.1) seriam os seguintes:

– ao nível de confiança de 90%, o escore padrão z = +/- 1,28

– o erro máximo admissível seria $zs_E = 1{,}28x5 = 6{,}4$

– então o intervalo onde cai o V seria $(100 - 6{,}4) \leq V \leq (100 + 6{,}4)$ ou

$$93{,}6 \leq V \leq 106{,}4$$

tendo uma amplitude de 12,8.

3.1.3 – Estimação via modelo da regressão linear

O modelo da regressão linear pretende predizer uma variável Y a partir de uma outra X e é expresso pela fórmula

$$Y' = \left(r_{XY} \frac{s_Y}{s_X} \right)(X - \overline{X}) + \overline{Y} \qquad (7.12)$$

Dedução:

Sejam os desvios y' = bx. Trata-se de estimar o valor de b que minimize os erros de estimação (y - y'). Já que E(y-y') = 0, o referido valor deverá minimizar os erros quadráticos, isto é, uma função desses erros: $f(E) = E(y-y')^2$.

Assim,

$$f(E) = E(y-bx)^2 = E(y^2) + b^2 E(x^2) - 2bE(xy)$$

$$= s_y^2 + b^2 s_x^2 - 2bCov(x,y)$$

$$= s_y^2 + b^2 s_x^2 - 2br_{xy} s_x s_y$$

Derivando f(E) para b, temos:

$$\frac{\delta f(E)}{\delta b} = 0 + 2b^2 s_x^2 - 2r_{xy} s_x s_y \text{ que igualando a zero dá}$$

$$0 = 2bs_x^2 - 2r_{xy} s_x s_y$$

$$b = \frac{2r_{xy} s_x s_y}{2s_x^2} = r_{xy} \frac{s_y}{s_x}$$

Consequentemente,

$$y' = r_{xy} \frac{s_y}{s_x} x \qquad \text{ou, em termos brutos,}$$

$$Y' - \overline{Y}' = r_{xy} \frac{s_Y}{s_X}(X - \overline{X}) \text{ então}$$

$$Y' = r_{XY} \frac{s_Y}{s_X}(X - \overline{X}) + \overline{Y}$$

Estimando o V a partir de T, a fórmula se apresenta como

$$V = \left(r_{TV} \frac{s_V}{s_T} \right) (T - \overline{T}) + \overline{V}$$

Mas como $\overline{V} = \overline{T} e \frac{s_V}{s_T} = r_{TV}$ (veja fórmula 7.3), segue que

$$V = (r_{TV} r_{TV})(T - \overline{T}) + \overline{T}$$
$$= r_{TV}^2 (T - \overline{T}) + \overline{T}$$

Mas, segundo a fórmula 7.3, $r_{TV}^2 = r_{tt}$; então

$$V = r_{tt} (T - \overline{T}) + \overline{T}.$$

Desta forma, no exemplo acima, onde $\overline{T} = 50$ e $r_{tt} = 0,75$

qual será o escore verdadeiro V de um escore empírico T=100?

Resposta: V = 0,75 (100 - 50) + 50 = 87,5.

A diferença entre o escore empírico T (= 100) e o predito V (= 87,5) é chamada de *erro de estimação*. Este é utilizado para estabelecer intervalos de confiança dentro dos quais se situa o escore verdadeiro. A utilização de intervalos em lugar de oferecer um escore individual único é uma praxe mais recomendada, dado que ela produz uma informação mais detalhada e precisa. Para estabelecer estes intervalos de confiança se utiliza o *erro padrão de estimação* (cf. capítulo 5) que vem dado pela fórmula

$$s_{VT} = s_E \sqrt{r_{tt}} \qquad (7.13)$$

Assim utilizando os dados do exemplo anterior, temos:

nível de confiança = 90%; N = 1.000; $s_E = 10$; $r_{tt} = 0,75$; T = 100; $\overline{T} = 50$.

Então

$$- s_{VT} = 10\sqrt{1 - 0,75}\sqrt{0,75} = 4,33$$

– erro máximo = $z \times s_{VX} = 1{,}28 \times 4{,}33 = 5{,}54$

– $V = r_{tt}(T - \overline{T}) + \overline{T} = 0{,}75(100 - 50) + 50 = 87{,}5$

– intervalo de confiança = V +/- erro máximo

$(87{,}5 - 5{,}54) \leq V \leq (87{,}5 + 5{,}54)$

$81{,}96 \leq V \leq 93{,}04.$

3.2 – Estimação da precisão das diferenças

É comum a aplicação de mais de um teste aos mesmos sujeitos para fins de seleção, treinamento, orientação acadêmica, encaminhamento psicológico, etc., que produzem tipicamente um perfil psicológico. Os resultados dos sujeitos nos diversos testes serão normalmente diferentes. Então se pergunta se estas diferenças são confiáveis ou como podem ser comparadas. Situação similar ocorre quando se quer comparar dois ou mais sujeitos num mesmo teste. Seus escores são comparáveis? Para efetivar estas tarefas existe o coeficiente de confiança das diferenças, pois comparar os resultados dos diferentes testes em termos dos respectivos desvios padrões é considerado um erro (Lord & Novick, 1968), a menos que os sujeitos ou os testes sejam escolhidos randomicamente, porque assim, no final das contas, isto é, depois de feitas infinitas seleções, as inferências baseadas nos desvios padrões teriam um sentido global.

Estas comparações são executadas, levando-se em conta o seguinte raciocínio: Para dois testes, A e B, a confiabilidade das diferenças entre seus escores (A – B = d) vem expressa por

$$r_{dd} = \frac{s_A^2 r_{AA} + s_B^2 r_{BB} - 2 s_A s_B r_{AB}}{s_A^2 + s_B^2 - 2 s_A s_B r_{AB}} \qquad (7.14)$$

onde,

r_{dd} : coeficiente de confiança das diferenças

s_A^2 e s_B^2 : variâncias dos escores do teste A e B

r_{AA} e r_{BB} : coeficiente de fidedignidade dos testes A e B

r_{AB} : correlação entre os dois testes.

Dedução:

$d = A - B$

O coeficiente de fidedignidade de d será

$$r_{dd} = \frac{s_V^2}{s_T^2} = \frac{E(V_A - V_B)^2}{E(A-B)^2} = \frac{E(V_A^2) + E(V_B^2) - 2E(V_A V_B)}{E(A^2) + E(B^2) - 2E(AB)} =$$

$$= \frac{s_{V_A}^2 + s_{V_B}^2 - 2\,\mathrm{cov}(V_A, V_B)}{s_A^2 + s_B^2 - 2\,\mathrm{cov}(A,B)}$$

Contudo, de 7.1 sabemos que $r_{tt} = \frac{s_{\bar{V}}}{s_T^2}$, donde

$s_{V_A}^2 = s_A^2 r_{AA}$ e $s_{V_B}^2 = s_B^2 r_{BB}$ e de 4.12 sabemos que $\mathrm{Cov}(V_A, V_B) = \mathrm{Cov}(A,B) = s_A s_B r_{AB}$.

Substituindo estas equivalências, temos

$$r_{dd} = \frac{s_A^2 r_{AA} + s_B^2 r_{BB} - 2 s_A s_B r_{AB}}{s_A^2 + s_B^2 - 2 s_A s_B r_{AB}}$$

Se os dois testes estiverem na mesma escala de medida (como, por exemplo, a escala padrão), as variâncias serão iguais ($s_A^2 = s_B^2$). Neste caso temos

$$r_{dd} = \frac{r_{AA} + r_{BB} - 2r_{AB}}{2(1 - r_{AB})} \qquad (7.15)$$

Dedução:

$s_A^2 = s_B^2$ e $s_A s_B = s_A^2$

Mas, de 7.15 evidenciando s_A^2 temos

$$r_{dd} = \frac{s_A^2(r_{AA} + r_{BB} - 2r_{AB})}{s_A^2(1+1-2r_{AB})}$$ que simplificando dá

$$r_{dd} = \frac{r_{AA} + r_{BB} - 2r_{AB}}{2(1-r_{AB})}$$

cujo erro padrão de medida das diferenças será

$$s_{ed} = s_d\sqrt{1-r_{dd}} \quad . \tag{7.16}$$

Exemplo:

Dois testes de inteligência, A e B, têm coeficientes de fidedignidade $r_{AA} = 0,80$ e $r_{BB} = 0,70$ e correlação $r_{AB} = 0,50$. Qual será o coeficiente de confiabilidade das diferenças dos escores dos sujeitos nos dois testes?

Resposta: $r_{dd} = \frac{0,80 + 0,70 - 2x0,50}{2(1-0,50)} = \frac{1,5-1}{1} = 0,50$ (fraco!).

III – Fatores que afetam a fidedignidade

Além das características dos próprios itens e do teste, há outros fatores, externos ao conteúdo do teste, que afetam a fidedignidade do mesmo. Dois destes fatores são particularmente relevantes, a saber: a variabilidade da amostra e o comprimento do teste. Aliás, esta é uma das instâncias importantes em que se evidenciam diferenças fundamentais entre a Psicometria Clássica e a Teoria da Resposta ao Item (IRT). Naquela, as características dos itens e do teste em sua totalidade dependem diretamente dos sujeitos (amostra) em que elas foram estabelecidas (*sample-dependent*) e um item depende dos outros itens do teste em sua caracterização individual, isto é, o item tem estas características e não outras (dificuldade, discriminação, etc.) porque está inserido neste conjunto de itens

(teste); se estivesse em outro conjunto manifestaria características diferentes (*test-dependent*).

1 – Variabilidade da amostra de sujeitos

Vimos que a equação da fidedignidade se baseia na correlação (entre testes paralelos). Agora, a correlação é afetada pelo tamanho da amostra de sujeitos utilizada para seu cômputo, pois quanto maior e variável a amostra de sujeitos, maior será o coeficiente de correlação e, consequentemente, o índice de fidedignidade. Disto segue que o coeficiente de fidedignidade de um teste não é fixo, mas varia segundo aumenta ou diminui a variabilidade da amostra de sujeitos. Aumentando a variabilidade da amostra, aumenta o índice de fidedignidade. Há fórmulas estatísticas para estimar o aumento deste índice com o aumento da variabilidade da amostra. Uma das fórmulas se baseia na variância erro que se supõe ser idêntica nas duas amostras (a de menor variabilidade e a de maior variabilidade) nas quais o índice de fidedignidade do teste é calculado. Assim, segundo a equação 7.4, temos

$$s_{E_1} = s_{T_1}\sqrt{1 - r_{11}}$$

$$s_{E_2} = s_{T_2}\sqrt{1 - r_{22}}$$

Mas, como $s_{E_1} = s_{E_2}$ segue que

$$s_{T_1}\sqrt{1 - r_{11}} = s_{T_2}\sqrt{1 - r_{22}}$$ que, elevando ao quadrado, dá

$$s^2_{T_1}(1 - r_{11}) = s^2_{T_2}(1 - r_{22}).$$

ou

$$s^2_{T_1}(1 - r_{11}) = s^2_{T_2} - s^2_{T_2} r_{22}.$$

Resolvendo para r_{22} dá

$$r_{22} = \frac{s^2_{T_2} - s^2_{T_1}(1 - r_{11})}{s^2_{T_2}}$$

$$= 1 - \frac{s^2_{T_1}}{s^2_{T_2}}(1 - r_{11}) \qquad (7.17)$$

Exemplo:

O coeficiente de fidedignidade de um teste de raciocínio verbal baseado em candidatos selecionados para um cargo foi de 0,70, e a variância foi de 30. Qual seria este coeficiente se tivesse sido calculado com todos os candidatos à seleção em vez de somente utilizar os selecionados, em cujo caso teríamos tido uma variância de 200?

Resposta:

$$r_{22} = 1 - \frac{30}{200}(1 - 0{,}70) = 0{,}96 .$$

Vê-se como a variabilidade da amostra afeta drasticamente o índice de fidedignidade de um teste que passou de 0,70 para 0,96 com o aumento da variância de 30 para 200. Note que esta fórmula é válida quando as variâncias erros das duas amostras forem iguais (Lord & Novick, 1968).

2 – Comprimento do teste

Também o número de itens do teste afeta a fidedignidade do mesmo. Quanto maior número de itens tiver o teste, maior será seu índice de precisão, pois o erro tende a zero quando o número se aproxima do infinito, segundo o famoso teorema de Bernoulli (cf. cap. 2).

O aumento da fidedignidade associado ao aumento do tamanho do teste é dado pela fórmula ou profecia de Spearman-Brown. Antes de apresentar esta fórmula, é importante notar que os itens que vão sendo acrescentados ao teste devem ser itens paralelos aos já presentes no teste, isto é, devem medir o mesmo traço latente. Óbvio!

A fórmula de Spearman-Brown é a seguinte:

$$r_{TT} = \frac{nr_{tt}}{1 + (n-1)r_{tt}} \qquad (7.18)$$

onde,

r_{TT}: fidedignidade do teste aumentado

r_{tt}: fidedignidade do teste original

n: número de vezes em que o teste original foi aumentado.

Por exemplo, no caso do cálculo do coeficiente de precisão pela técnica das duas metades, o teste original (isto é, a metade 1) foi aumentado duas vezes, a saber, metade 1 + metade 2. Neste caso, a fórmula de Spearman-Brown será:

$$r_{TT} = \frac{2r_{tt}}{1 + r_{tt}}$$

Exemplo:

Um teste de 30 itens foi aplicado a uma amostra de sujeitos e se obteve um coeficiente de fidedignidade de 0,80. Qual seria este coeficiente se aos 30 itens fossem acrescidos mais 20 itens paralelos?

Resposta:

$n = \frac{30 + 20}{30} = 1{,}67$, o teste foi aumentado 1,67 vezes. Então,

$$r_{TT} = \frac{1{,}67 \text{x} 0{,}80}{1 + (1{,}67 - 1)\text{x}0{,}80} = 0{,}87$$

Assim, aumentando o tamanho do teste de 30 para 50 itens, o teste ganha em precisão, passando esta de 0,80 para 0,87.

Note, entretanto, que a profecia de Spearman-Brown também prediz a queda do coeficiente de fidedignidade se forem eliminados itens de um teste.

Exemplo:

Um teste de 200 itens teve índice de fidedignidade de 0,96. Se tirarmos 100 itens, qual será o novo índice de precisão?

Resposta:

$n = \frac{200 - 100}{200} = 0{,}50$ o teste foi reduzido pela metade. Então,

$$r_{TT} = \frac{0{,}50 \times 0{,}96}{1 + (0{,}50 - 1) \times 0{,}96} = 0{,}92 .$$

Assim, a retirada de 100 itens dos 200 originais reduziu o índice de fidedignidade do teste de 0,96 para 0,92.

Pode-se igualmente querer aumentar o coeficiente de fidedignidade de um teste para um valor desejado. Por exemplo, um teste de 40 itens tem coeficiente de fidedignidade de 0,75. Eu quero um coeficiente mais aceitável, de pelo menos 0,90. Quantos itens tenho que acrescentar aos 40 existentes no meu teste para conseguir tal índice?

Resposta:

Da fórmula 7.18, podemos descobrir a fórmula para o cálculo do *n*, que é a seguinte:

$$n = \frac{r_{TT(1-r_{tt})}}{r_{tt(1-r_{TT})}} \qquad (7.19)$$

Assim, temos que $r_{TT} = 0{,}90$ e $r_{tt} = 0{,}75$

Então,

$$n = \frac{0{,}90(1 - 0{,}75)}{0{,}75(1 - 0{,}90)} = 3$$

É preciso triplicar o número de itens do teste (3 x 40 = 120) para poder passar de um coeficiente de fidedignidade de 0,75 para 0,90; isto implica em ter que acrescentar 80 novos itens paralelos (120 - 40 = 80).

Dedução da fórmula 7.19:

$$r_{TT} = \frac{nr_{tt}}{1+(n-1)r_{tt}}$$ Disto segue que

$$nr_{tt} = r_{TT}\left[1+(n-1)r_{tt}\right]$$

$$= r_{TT} + (n-1)r_{tt}r_{TT}$$

$$= r_{TT} + nr_{tt}r_{TT} - r_{tt}r_{TT}$$

Enviando os termos com n para a esquerda e efetuando, temos

$$nr_{tt} - nr_{tt}r_{TT} = r_{TT} - r_{tt}r_{TT}$$

$$n\left(r_{tt} - r_{tt}r_{TT}\right) = r_{TT} - r_{tt}r_{TT}$$

$$n = \frac{r_{TT} - r_{tt}r_{TT}}{r_{tt} - r_{tt}r_{TT}}$$

$$= \frac{r_{TT}\left(1-r_{tt}\right)}{r_{tt}\left(1-r_{TT}\right)}$$

Concluindo este capítulo sobre a fidedignidade dos testes, o leitor deve ter-se dado conta que se trata do capítulo onde a parafernália estatística se apresenta mais complexa e sofisticada em Psicometria. Duas observações: primeiro, tal sofisticação se coaduna com a Psicometria Clássica, pois a fidedignidade diz respeito à calibração dos instrumentos, isto é, à precisão da medida de eventos empíricos (no caso, os comportamentos, os itens), onde a visão positivista funciona bem; mas, em segundo lugar, todos estes índices da parafernália estatística na estimação da fidedignidade produzem praticamente os mesmos resultados e, portanto, a parafernália aparece com um matiz de curiosidade acadêmica ou jogo estatístico. Isto implica que qualquer coeficiente que você queira usar, entre os apresentados e outras dezenas deles aqui não mencionados, produz a estimação da fidedignidade dos testes que você deseja.

CAPÍTULO 8
Normatização dos testes

Introdução

Padronização ou normatização, em seu sentido mais geral, refere-se à necessidade de existir uniformidade em todos os procedimentos no uso de um teste válido e preciso: desde as precauções a serem tomadas na aplicação do teste (uniformidade das condições de testagem, controle do grupo, instruções padronizadas e motivar os examinandos pela redução da ansiedade) até o desenvolvimento de parâmetros ou critérios para a interpretação dos resultados obtidos. Em seu sentido mais técnico de parâmetro psicométrico, a padronização se refere a este último aspecto, isto é, como interpretar os resultados obtidos num teste. Alguns autores (incluindo Cronbach, 1996) querem fazer uma distinção clara entre

- Padronização, como sendo a uniformidade na aplicação dos testes e
- Normatização, sendo a uniformidade na interpretação dos escores dos testes.

A distinção é importante, porque fala de duas questões muito distintas. Entretanto, a literatura neste particular não é consistente com a nomenclatura; pelo contrário, as duas expressões são utilizadas indistintamente. Contudo, como se trata de questões distintas, vamos tratar o tema em duas seções separadas neste capítulo.

I – Padronização das Condições de Administração dos Testes Psicológicos

A padronização das condições de aplicação dos testes psicológicos tem como preocupação garantir que a coleta dos dados sobre os sujeitos seja de boa qualidade. De fato, uma má aplicação torna os dados obtidos

inválidos, mesmo quando obtidos através de um teste de boa qualidade. A má aplicação não invalida a qualidade, digamos psicométrica, do teste (se ele é um teste válido e preciso, ele continua sendo assim), mas torna o protocolo do sujeito inválido, isto é, os dados obtidos sobre este sujeito não são confiáveis. Assim, uma má aplicação do teste estraga a utilidade do mesmo, pelo mau uso que dele se faz. Então, a padronização das condições de testagem pretende garantir o uso adequado e legítimo dos testes psicológicos. Claro que tal preocupação é relevante e importante somente se o teste ele mesmo for de boa qualidade; uma boa aplicação de um teste inválido não salva nada, os resultados continuam inválidos.

Para se garantir uma boa administração dos testes psicológicos é preciso atender a requisitos referentes aos seguintes temas:

– O material da testagem

– A aplicação dos testes (o ambiente da testagem)

1 – O material de testagem

Quanto ao material da testagem, duas condições devem ser atendidas:

a) *Qualidade do teste*: o teste tem que ser válido e preciso, como foi definido nos capítulos 6 e 7; o uso de testes sem estes parâmetros é inútil, eticamente condenável e judicialmente processável. Na verdade, o uso de testes sem tais parâmetros qualifica o seu usuário como charlatão, terminologia que define o usuário como criminoso diante da lei, e como eticamente irresponsável diante do conhecimento científico;

b) *Pertinência do teste*: além de ser válido e preciso, o teste deve: (1) ter relevância ao problema apresentado pelo sujeito testando. Nenhum teste serve para toda e qualquer avaliação. O aplicador deve saber para que serve um dado teste e escolher aquele que se aplica ao problema do testando. Este é um problema bastante grave para o psicólogo, uma vez que, apesar de haver tantos testes no mercado, não existem testes para todas as necessidades que os sujeitos podem apresentar. Veja, por exemplo, o caso da seleção: praticamente quase não existem testes construídos para este fim; assim, na hora de escolher os testes para tal intento, o psicólogo tem que se virar para utilizar

testes que, pelo menos, possam dar alguma informação pertinente para tal e tal cargo. Assim, por exemplo, um teste de raciocínio dedutivo dificilmente se justifica num psicotécnico para motoristas amadores. Além de ser pertinente ao caso, (2) o teste escolhido deve se adaptar ao nível do candidato, isto é, adaptado ao nível intelectual, profissional, etc. do candidato. Por exemplo, testes de tipo verbal, onde se exige leitura e compreensão, não são pertinentes no caso de testagem de analfabetos ou de crianças.

2 – Aplicação dos testes psicológicos (O ambiente de testagem)

Os testes são instrumentos técnicos e seu manejo geralmente necessita de pessoal treinado e conhecedor, como qualquer aparelho de tecnologia sofisticada como são os testes. Assim, nem todo o mundo é capaz ou pode aplicar testes psicológicos, além de serem inclusive de uso exclusivo da profissão dos psicólogos. Sendo instrumentos sofisticados, os testes requerem uma série de regras para sua aplicação, regras que são expressas sob o que se chama de *padronização* da aplicação dos mesmos. O que é que implica tal padronização? Ela implica em várias coisas, particularmente na observância de

– procedimentos de aplicação

– direito dos testandos

– controle dos vieses do aplicador

– normas na divulgação dos resultados.

2.1 – Administração dos testes

Os procedimentos na aplicação dos testes têm como objetivo garantir a validade da testagem, porque um teste técnica e cientificamente válido e pertinente ao caso pode produzir resultados inválidos se for mal aplicado. O que é que tornaria uma testagem inválida pela má aplicação? Uma resposta genérica seria a seguinte: os resultados do teste são válidos (obviamente supondo que o próprio teste seja válido) se a sua aplicação seguiu à risca as instruções e recomendações dadas pelo seu autor, isto é, se o aplicador seguiu exatamente o manual de aplicação do teste. Nor-

malmente, tais orientações irão exigir pelo menos duas condições de aplicação para que os resultados sejam válidos e confiáveis, a saber,

– a qualidade do ambiente físico da aplicação

– a qualidade do ambiente psicológico, que tipicamente significa uma atmosfera em que a ansiedade do testando seja reduzida ao mínimo.

Quanto ao *ambiente físico*: todas as condições do ambiente físico devem ser tais que ponha o testando em condições ótimas de ação. Como se quer saber, com o teste, o nível de aptidão ou as preferências do testando, este deve se sentir na sua melhor forma para poder agir exatamente de acordo com suas habilidades, interesses e pendores e não influenciado por fatores estranhos oriundos do meio ambiente. Assim, se necessita que o meio ambiente não produza distratores em termos fisiológicos e psicológicos para o testando. Desta forma, é preciso tomar cuidado com

– posto de trabalho: cadeira, mesa, espaço físico

– condições atmosféricas: iluminação, temperatura, ventilação, higiene

– condições de silêncio: isolamento acústico, ausência de interrupções

– apresentação do aplicador: roupas limpas e adequadas, vocabulário apropriado, uso de perfumes

– evitar interrupções durante a testagem.

Quanto às *condições psicológicas*, é preciso atender a que

– o testando esteja em condições normais de saúde física e psicológica; no caso de diagnóstico psiquiátrico, o sujeito deve querer se submeter a testagem ou esta tenha sido encomendada pelo responsável do paciente;

– o testando compreenda exatamente a tarefa a executar: isto pode implicar que as instruções do teste devam ou não ser dadas em voz alta; o aplicador deve responder a todas as questões referentes à compreensão da tarefa, sem dar dicas de solução para as próprias questões do teste. Esta tarefa evidentemente é mais delicada na testagem de crianças e pessoas com outras dificuldades (deficientes, surdos-mudos, etc.). De qualquer forma,

aqui é preciso atender a duas coisas: primeiro, o sujeito deve entender perfeitamente a tarefa que é dele pedida (assim, às vezes são necessárias explicações ulteriores) e, segundo, mudar as instruções do manual implica ou pode implicar em mudar o próprio teste (isto é, invalida o teste). De sorte que, no ideal, as instruções devem ser dadas uma única vez e iguais para todo o mundo. Tudo isso importa em que o aplicador seja profundamente familiarizado com o teste;

– o nível de ansiedade do testando seja reduzido: isto implica no estabelecimento do *rapport*. Este é mais importante na testagem individual. Em que ele consiste? Existe o rapport quando o testando vê no aplicador um amigo e não um estranho e menos ainda um carrasco. Significa, no fundo, que o testando se sinta bem à vontade ao fazer o teste, o que implica que o examinador seja motivador e encorajador, não se irrite, não grite ou faça cara feia, etc.

Agora, como proceder em *situações adversas*? Situações adversas de testagem são, por exemplo, a aplicação de testes para fins periciais e a testagem para seleção. Nestes casos, o sujeito se encontra necessariamente em condições psicológicas e às vezes até físicas, não satisfatórias, sobretudo porque ele está ali como vítima ou numa situação de alta competição. Onde ficou, então, o estado ideal psicológico da pessoa para poder tomar o teste de uma maneira adequada? No caso da testagem de seleção, especialmente em concursos públicos, parece até haver uma contradição entre a situação ideal de aplicação dos testes e a exigência constitucional da isonomia, segundo a qual todo o mundo deve ser tratado identicamente. Quando você tem milhares de candidatos concorrendo para um cargo, uma testagem individual é proibitivamente difícil de ser realizada; então, tipicamente se faz testagem em grupo. Neste caso, todo o mundo deve ser tratado do mesmo modo, o que implica mesmo horário, mesmas condições, mesmo tudo. Mas se um sujeito está doente ou de qualquer forma impossibilitado de tomar adequadamente o teste naquele dia, que fazer? Remarcar para outro dia pode estar ofendendo o princípio legal da isonomia, ao dar a ele tratamento diferenciado; tomar o teste naquelas condições adversas pode produzir resultados inválidos... Na prática, o que tem prevalecido em tais situações são os ditames da isonomia.

No presente, tais situações de testagem são uma dor de cabeça e angústia para o psicólogo responsável e consciente de sua profissão.

2.2 – Comportamento e vieses do examinador

Na situação de testagem, em especial a testagem individual, o aplicador do teste é um elemento importante da situação. Seu modo de ser e de atuar podem afetar bastante os resultados do teste. As pesquisas que existem sobre este assunto geralmente não permitem conclusões decisivas sobre o grau de influência que estas variáveis do examinador têm sobre os resultados dos testes. De qualquer forma, seriam importantes as seguintes situações:

– O examinador deve ser familiar ao testando?

– Encorajar frequentemente o testando ajuda ou atrapalha?

– O sexo do examinador é relevante?

– A idade do examinador é relevante?

– O estado emocional do examinador é relevante?

– As atitudes e opiniões pessoais do examinador são relevantes?

Enfim são muitas perguntas para se poder dar uma resposta sensata. As pesquisas, como disse, nem sempre dão respostas unânimes sobre tais questões, mas simplesmente ignorar tais questões seria supor que elas não têm relevância nas relações sociais, e a situação de testagem é uma relação social. Pelo menos, o psicólogo deve estar consciente da possível influência de tais fatores e procurar minimizá-la. Inclusive, dizem as más línguas que um dos motivos mais fortes de porque alguém se torna psicólogo é porque ele ou ela quer resolver seus próprios problemas psicológicos pessoais e, ao exercer sua profissão, o psicólogo entra com todos estes problemas, deixando escapar dicas sutis, mas eficazes, que influenciam negativa ou positivamente o comportamento dos outros, dos testandos no caso. Por exemplo, se o examinador instintivamente gosta ou não gosta do testando, não irá ele sinalizar, de alguma forma, através do seu comportamento, tal sentimento? E se o fizer, o testando não será prejudicado em seu desempenho no teste? Se o testando sai de lá com a impressão de que o psicólogo foi antipático, rude, etc., então certamente foi negativamente influenciado pelo examinador. O psicólogo é um ser humano como todos os outros, com seus problemas inclusive, mas ele é

também um técnico ou perito que deve ter desenvolvido algumas habilidades próprias da profissão, das quais obviamente ele deve fazer uso em situações como a testagem psicológica. A primeira delas, quiçá, seria a de um autoconhecimento mais elaborado que lhe permita conhecer melhor seus fortes e fracos e, assim, poder lidar com eles na sua profissão.

De qualquer forma, o aplicador dos testes psicológicos, o qual tipicamente deve ser um psicólogo, deve atender aos seguintes requisitos:

1) *Conhecimento*: o aplicador deve conhecer profundamente o material utilizado, para que possa oferecer respostas às questões levantadas pelos candidatos e transmitir-lhes segurança;

2) *Aparência*: trata-se de usar roupas adequadas e limpas, pois o aplicador deve causar boa impressão; enfim, ele deve se apresentar como um perito, evitando extravagâncias na sua apresentação; utilizar perfumes não extravagantes, ...;

3) *Comportamento durante a testagem*: o aplicador está aí para conduzir a testagem, assim, ele deve manter ordem, respeito, orientação, sem fazer interferências e interrupções desnecessárias. Assim, é importante atender ao uso da linguagem, utilizando vocabulário adequado e compreensível ao grupo ou sujeito. Como ele deve transmitir seriedade, segurança e confiança aos testandos, suas atitudes devem ser correspondentes: ser atencioso, não se irritar ou gritar, movimentar-se discretamente na sala (utilizar sapatos que não façam barulho).

2.3 – O direito dos testandos

Na sociedade democrática moderna, este tema se tornou algo de importância fundamental, isto é, os direitos das pessoas, garantidos até nas normas da Constituição dos países e das Nações Unidas. No Brasil, inclusive, a atuação do psicólogo na testagem é considerada uma atividade pericial, isto é, atividade de perito. Por lei, os peritos devem prestar serviços de qualidade à sociedade e esta qualidade pode ser judicialmente procurada através das leis pertinentes, como as do PROCON, por exemplo. De tal forma que o psicólogo responde até criminalmente por sua conduta nesta área dos testes. A lei considera o psicólogo como perito e, portanto, legalmente responsável em sua atuação de profissional.

O Prof. Norberto Abreu e Silva Neto (Santos & Abreu e Silva Neto, 2000) dá as seguintes informações sobre esta questão:

"**O princípio do consentimento livre e esclarecido**

Em 1947, após o Tribunal Internacional de Nuremberg ter definido os crimes contra a humanidade, um tribunal americano encarregado de julgar os atos de "barbárie científica" cometidos pelos médicos nazistas acabou por estabelecer o Código de Nuremberg, tendo em vista garantir fossem respeitadas as pessoas humanas que viessem a participar de experimentos médicos ou científicos. O princípio essencial estabelecido por esse Código é de que toda experimentação com seres humanos requer o prévio consentimento livre e esclarecido do sujeito participante. Assim, com base nesse princípio são feitas recomendações práticas que o complementam e que servem como normas éticas para a realização de experimentos científicos.

No ano seguinte, surge um documento fundamental ao movimento ético. Por considerar que as democracias anteriores à Segunda Grande Guerra haviam sido complacentes com as ditaduras e que estas, por seu lado, acabavam por colocar em perigo a paz internacional, a Comunidade Internacional dos Juristas proclama a Declaração Universal dos Direitos do Homem, destinada a promover sua proteção contra os regimes ditatoriais e que foi adotada pela Assembleia Geral das Nações Unidas, em 10 de dezembro de 1948.

Outro documento importante desse período inicial do movimento ético é o Manifesto Russell-Einstein, lançado em 1955, em resposta às monstruosidades das duas guerras e também em função das novas ameaças trazidas pela "Guerra Fria". Esse Manifesto nos é apresentado por Toulouse (1998: 43) como *o primeiro reconhecimento solene pelos cientistas de uma responsabilidade coletiva* pelo impacto social da ciência. Ao Manifesto segue-se o estabelecimento da Conferência Pugwash para as Ciências e Questões Mundiais, fórum da comunidade científica internacional para a promoção ativa dos valores universais de racionalidade e objetividade e que se propõe, diz Toulouse, *estar ao lado e acima das divisões políticas e ideológicas* (p. 34).

Todavia, apesar desse esforço ético, sabemos por experiência própria que as ditaduras e as guerras não deixaram de existir nos últimos quarenta anos. De modo semelhante, o Código de Nuremberg mostrou-se logo insuficiente para garantir o bem-estar dos sujeitos de pesquisas contra os cientistas desejosos de uma ciência sem fronteiras ou limites. Assim, surge, em 1964, a Declaração de Hel-

sinque, adotada pela 18ª Assembleia Médica Mundial e depois emendada nas Assembleias de 1975 (Tóquio) e de 1983 (Veneza), e que contém recomendações destinadas a guiar os médicos nas pesquisas biomédicas.

O bem-estar dos sujeitos da pesquisa como prioridade e os Comitês de Ética

A Declaração de Helsinque reafirma o princípio do consentimento livre e esclarecido e coloca o bem-estar do sujeito como prioritário quando afirma: *Na pesquisa médica os interesses da ciência e aqueles da sociedade não devem jamais prevalecer sobre o bem-estar dos sujeitos* (apud Ambroselli, 1987: 8). Mas, a maior inovação trazida por essa declaração, informa-nos Ambroselli, é a proposta do estabelecimento dos Comitês de Ética, uma tentativa da comunidade científica no sentido de *lutar contra certos experimentos escandalosos em execução, notadamente nos Estados Unidos* (p. 6). Assim, a partir dessa data, todos os projetos de pesquisa na área biomédica envolvendo seres humanos deverão ser "revisados" por Comitês de Ética e de acordo com as normas estabelecidas na Declaração. Essa recomendação se estende à publicação dos resultados de pesquisas, que, como os projetos, deverão ser "revisados" eticamente.

De acordo com Ambroselli (1987), a Declaração de Helsinque exerceu uma dupla influência sobre o movimento ético: primeiro, diz, *pela instalação dos comitês de ética, notadamente nos Estados Unidos, encarregados de fazer cumprir essas diretivas pelos promotores de pesquisa ou pelas revistas científicas que publicavam essas pesquisas* (p. 6-7). Em segundo lugar, ela acrescenta, sua influência deu-se de modo indireto. Por pressão de financiadores de pesquisa norte-americanos, outros países começaram a desenvolver comitês de ética.

No final dos anos sessenta e durante os setenta surgem novas preocupações e contestações da sociedade quanto aos efeitos nocivos do "progresso científico" sobre a saúde humana e do planeta: os efeitos de radiações, as diversas formas de poluição industrial, a perspectiva de esgotamento dos recursos naturais, a explosão demográfica etc. Outra fonte de inquietações provinha, nessa época, da emergência de avanços técnicos em biomedicina que davam origem a fenômenos espantosos: transplante de órgãos, diagnóstico pré-natal, fecundação *in vitro*, definição de morte legal etc. Assim, em 1975, biólogos-geneticistas, reunidos em Conferência por iniciativa da *Royal Society* britânica e da NAS americana, decretaram a "moratória de Asilomar" referente aos organismos geneticamente modificados. De acordo com Ambroselli

(1987), essa moratória fez surgir comissões nacionais de segurança e comitês de ética criados por instâncias governamentais para fazer frente ao *perigo potencial das moléculas de DNA recombinadas* (p. 9) denunciado no próprio título da Conferência de Asilomar".

Com base nestes princípios, os comitês de ética em Psicologia, inclusive no Brasil, vêm elaborando normas que devem ser seguidas na aplicação de testes. De um modo geral estas normas podem ser resumidas segundo as Normas para a Testagem Educacional e Psicológica da American Psychological Association (APA, 1985, apud Cronbach, 1996, p. 97) nos seguintes pontos:

1) Deve ser obtido o consentimento informado dos testandos ou de seus representantes legais antes da realização da testagem. As exceções a esta regra são: a) testagem por determinação legal (perícia) ou governamental (testagem nacional); b) testagem como parte de atividades escolares regulares; c) testagem de seleção, onde a participação implica em consentimento;

2) Em aplicações escolares, clínicas e de aconselhamento, os testandos têm direito a explicações em linguagem que eles compreendem com respeito aos resultados que os testes irão produzir e das recomendações que deles decorrem;

3) Em aplicações escolares, clínicas e de aconselhamento, quando os escores são utilizados para tomar decisões que afetam os testandos, estes ou seus representantes legais têm o direito de conhecer seu escore e a sua interpretação.

2.4 – Sigilo e divulgação dos resultados

Com respeito à divulgação do resultado dos testes, devem ser seguidas as normas do sigilo profissional. Os seguintes pontos esclarecem alguns dos princípios a serem seguidos pelo psicólogo profissional:

1) *Quem tem direito aos resultados* dos testes? Certamente o *candidato* que se submeteu aos testes tem o direito a toda e qualquer informação que desejar. Concomitante a este direito a todo o conteúdo dos testes, o psicólogo deve respeitar também o princípio do consentimento *informado* do candidato, o qual

lhe dá o direito a que o psicólogo utilize uma linguagem acessível ao candidato com referência a qualquer informação relevante que os testes produzem. Também tem direito aos resultados o *solicitante* da testagem, como o dono da empresa no caso da seleção ou o juiz no caso da perícia judicial. O direito destes, entretanto, não é sobre todos os resultados obtidos na testagem; eles apenas têm direito às informações estritamente necessárias à resposta da solicitação. Assim, se a solicitação foi a de que o psicotécnico indicasse se o candidato foi apto ou não para o cargo em disputa, esta é a única informação relevante e necessária para o empregador; todo o resto de informação que os testes produzem é privilégio individual do candidato e somente ele tem direito a tais informações;

2) O *sigilo* e a *segurança* dos resultados dos testes devem, em geral, seguir as normas do sigilo entre profissional e paciente, similarmente ao sigilo médico. Assim, (a) os arquivos devem ser seguros de modo que ninguém possa ter acesso a um dado caso sem autorização específica do profissional responsável; (b) as identidades dos indivíduos devem ser codificadas de tal forma que somente o profissional responsável seja capaz de identificar; (c) em processos judiciais, o juiz pode pedir abertura de registros sigilosos. Como proceder em tais casos? Este problema não tem ainda solução satisfatória na prática e é causa de diatribes que normalmente são resolvidas em tribunal de justiça. Normalmente, também, o juiz quer sempre mais do que o psicólogo está disposto a oferecer, amparando-se este no sigilo profissional... De qualquer forma, o indivíduo não pode sair *indevidamente prejudicado* com a exposição de informações sigilosas;

3) O Código de Ética do Psicólogo no Brasil (CFP, 1987) diz o seguinte sobre esta questão:

Art. 02 – Ao Psicólogo é vedado:

b) Apresentar, publicamente, através dos meios de comunicação, resultados de psicodiagnóstico de indivíduos ou grupos,

bem como interpretar ou diagnosticar situações problemáticas, oferecendo soluções conclusivas;

l) Interferir na fidedignidade de resultados de instrumentos e técnicas psicológicas;

m) Adulterar resultados, fazer declarações falsas e dar atestado sem a devida fundamentação técnico-científica.

Art. 03 – São deveres do psicólogo nas suas relações com a pessoa atendida:

a) Dar à(s) pessoa(s) atendida(s) ou, no caso de incapacidade desta(s), a quem de direito, informações concernentes ao trabalho a ser realizado;

b) Transmitir a quem de direito somente informações que sirvam de subsídios às decisões que envolvem a pessoa atendida;

c) Em seus atendimentos, garantir condições ambientais adequadas à segurança da(s) pessoa(s) atendida(s), bem como à privacidade que garanta o sigilo profissional.

Art. 06 [O Psicólogo em Instituições Empregadoras] – O Psicólogo garantirá o caráter confidencial das informações que vier a receber em razão de seu trabalho, bem como do material psicológico produzido.

Parágrafo 1 – Em caso de demissão ou exoneração, o Psicólogo deverá repassar todo o material ao psicólogo que vier a substituí-lo.

Parágrafo 2 – Na impossibilidade de fazê-lo, o material deverá ser lacrado na presença de um representante do CRP, para somente vir a ser utilizado pelo Psicólogo substituto, quando, então, será rompido o lacre, também na presença de um representante do CRP.

Parágrafo 3 – Em caso de extinção do serviço psicológico, os arquivos serão incinerados pelo profissional responsável, até aquela data, por este serviço, na presença de um representante do CRP.

Art. 19 [O Psicólogo e a Justiça] – Nas perícias, o Psicólogo agirá com absoluta isenção, limitando-se à exposição do que tiver conhecimento através do seu trabalho e não ultrapassando, nos laudos, o limite das informações necessárias à tomada de decisão.

Do Sigilo Profissional

Art. 21 – O sigilo protegerá o atendido em tudo aquilo que o Psicólogo ouve, vê ou de que tem conhecimento como decorrência do exercício da atividade profissional.

Art. 22 – Somente o examinado poderá ser informado dos resultados dos exames, salvo nos casos previstos neste Código.

Art. 23 – Se o atendimento for realizado por Psicólogo vinculado a trabalho multiprofissional numa clínica, empresa ou instituição ou a pedido de outrem, só poderão ser dadas informações a quem as solicitou, a critério do profissional, dentro dos limites do estritamente necessário aos fins a que se destinou o exame.

Parágrafo 1 – Nos casos de perícia, o Psicólogo tomará todas as precauções, a fim de que só venha a relatar o que seja devido e necessário ao esclarecimento do caso.

Parágrafo 2 – O Psicólogo, quando solicitado pelo examinado, está obrigado a fornecer a este as informações que foram encaminhadas ao solicitante e a orientá-lo em função dos resultados obtidos.

Art. 24 – O Psicólogo não remeterá informações confidenciais a pessoas ou entidades que não estejam obrigadas ao sigilo por Código de Ética ou que, por qualquer forma, permitam a estranhos o acesso a essas informações.

II – Normatização dos Testes Psicológicos

A normatização diz respeito a padrões de como se deve interpretar um escore que o sujeito recebeu num teste. Isto porque um escore bruto produzido por um teste necessita ser contextualizado para poder ser interpretado. Obter, por exemplo, 50 pontos num teste de raciocínio verbal e 40 num de personalidade não oferece nenhuma informação. Mesmo se

dissermos que ele acertou 80% das questões não diz muito, visto que o teste pode ser fácil (80% então seria pouco) ou difícil (80% então seria muito). Na verdade, qualquer escore deve ser referido a algum padrão ou norma para adquirir sentido. Uma tal norma permite situar o escore de um sujeito, permitindo: (1) determinar a posição que o sujeito ocupa no traço medido pelo teste que produziu o tal escore e, (2) comparar o escore deste sujeito com o escore de qualquer outro sujeito.

O critério de referência de que estamos falando ou a norma de interpretação é constituído tipicamente por três padrões: (1) o nível de desenvolvimento do indivíduo humano (normas de desenvolvimento), (2) um grupo padrão constituído pela população típica para a qual o teste é construído (normas intragrupo) e, (3) um critério externo (normas referentes a critério – *criterion-referenced norms*), este utilizado particularmente no caso de testes de aprendizagem.

1 – Normas de desenvolvimento

As normas de interpretação dos escores de um teste baseadas no desenvolvimento se fundamentam no fato do desenvolvimento progressivo (nos vários aspectos de maturação psicomotora, psíquica, etc.) pelo qual o indivíduo humano passa ao longo de sua vida. Neste sentido, são utilizados, como critério de norma, três fatores, a saber: idade mental, série escolar, estágio de desenvolvimento.

1.1 – A idade mental

A idade mental como critério foi criado por Binet e Simon (1905). Estes autores falavam de nível mental, depois popularizado como idade mental. Binet e Simon separaram empiricamente uma série de 54 questões/tarefas em 11 níveis de idade cronológica: 3 a 10 anos (oito níveis), 12, 15 anos e idade adulta. As questões que eram respondidas corretamente pela média de crianças/sujeitos de uma idade cronológica X definiam o nível/idade mental correspondente a esta idade cronológica. Assim, a um sujeito que respondia a todas as questões que as crianças de 10 anos eram capazes de responder era atribuída a idade mental de 10 anos.

Na adaptação norte-americana da escala de Binet-Simon, a Stanford-Binet (Terman & Merrill, 1960), a idade mental (IM) foi expressa

em termos da idade cronológica (IC), resultando no quociente intelectual, o QI, através da fórmula:

$$QI = 100 \times \frac{IM}{IC}.$$

Assim, quem responde a todas as questões correspondentes à sua idade cronológica possui um QI de 100 (por exemplo, para uma criança de 10 anos: QI = 100 x (10/10) = 100). A interpretação dos resultados em termos de QI se fazia através das categorizações apresentadas na tabela 8-1.

Tabela 8-1. Interpretação dos escores de QI

QI	Interpretação
140 – 160	Definitivamente Superior
120 – 139	Superior
110 – 119	Médio superior
90 – 109	Normal ou médio
80 – 89	Médio inferior
70 – 79	Deficiência-limítrofe
50 – 69	Cretino
30 – 49	Imbecil
– 29	Idiota

Note que o QI de que se fala hoje em dia não é este QI, mas ele é uma transformação das normas baseadas em *z*, como veremos mais adiante.

1.2 – Série escolar

Este critério é utilizado para testes de desempenho acadêmico e somente faz sentido quando se trata de disciplinas que são oferecidas numa sequência de várias séries escolares. As normas são aqui estabelecidas, computando-se o escore bruto médio obtido pelos alunos em cada série, resultando

num escore típico para cada série. Desta forma, a criança que obtém o escore bruto típico da 4ª série recebe o escore padronizado de 4.

1.3 – Estágio de desenvolvimento

Este critério é utilizado por pesquisadores na área da psicologia da criança que estudam o desenvolvimento mental e psicomotor em termos de idades sucessivas de desenvolvimento, como Gesell e Piaget.

Gesell e colaboradores (Ames, 1937; Gesell & Amatruda, 1947; Halverson, 1933; Knoblock & Pasamanick, 1974) desenvolveram normas para oito idades típicas (de 4 semanas a 36 meses) de desenvolvimento das crianças nas áreas do comportamento motor, adaptativo, da linguagem e social.

Piaget e seus colaboradores estudaram o desenvolvimento cognitivo e estabeleceram uma sequência de estágios sucessivos deste desenvolvimento (sensório-motor, pré-operacional, operacional concreto, operacional formal). Seguidores da escola piagetiana desenvolvem testes utilizando estes estágios como critério de interpretação dos escores (Laurendeau & Pinard, 1962, 1970; Pinard & Laurendeau, 1964).

2 – Normas intragrupo

Nas normas intragrupo, o critério de referência dos escores é o grupo ou a população para a qual o teste foi construído. Aqui o escore do sujeito toma sentido em relação aos escores de todos os sujeitos da população. De fato, ele é referenciado em termos (1) do posto percentílico ou (2) do desvio padrão (z). Como tipicamente não são conhecidos os escores da população, é sobre uma amostra representativa dela que são estabelecidas as normas.

2.1 – Posto percentílico

O escore do sujeito é expresso em termos de percentil. Este posto indica quanto por cento de todos os sujeitos da população (amostra) estão abaixo do escore do sujeito. Assim, se 40% dos sujeitos obtiveram um escore bruto menor do que 20, este escore será expresso como percentil 40, o que indica que 40% dos sujeitos têm escore menor que 20 e 60% têm escore maior. Um percentil de 50 indica que o sujeito se situa na me-

diana dos escores da amostra. Usa o intervalo semi-interquartílico (Q) em torno desta mediana para definir o significado relativo dos postos dos sujeitos. Q = (Q3 – Q1)/2, onde Q1 é o percentil 25 e o Q3 o percentil 75.

Os escores percentílicos são fáceis de calcular e são de compreensão simples. A grande dificuldade da escala percentílica consiste no fato dela ser uma escala ordinal, isto é, as distâncias entre escores sucessivos não são constantes, mas variam segundo a posição do escore estar no início/fim da escala ou no meio dela. De fato, os intervalos entre os percentis (PR) medianos são menores do que os dos extremos da escala, como aparece ilustrado na figura 8-1. Isto implica em que diferenças iguais entre percentis não significam diferenças iguais em termos dos escores brutos. De fato, os percentis constituem uma transformação não linear dos escores brutos.

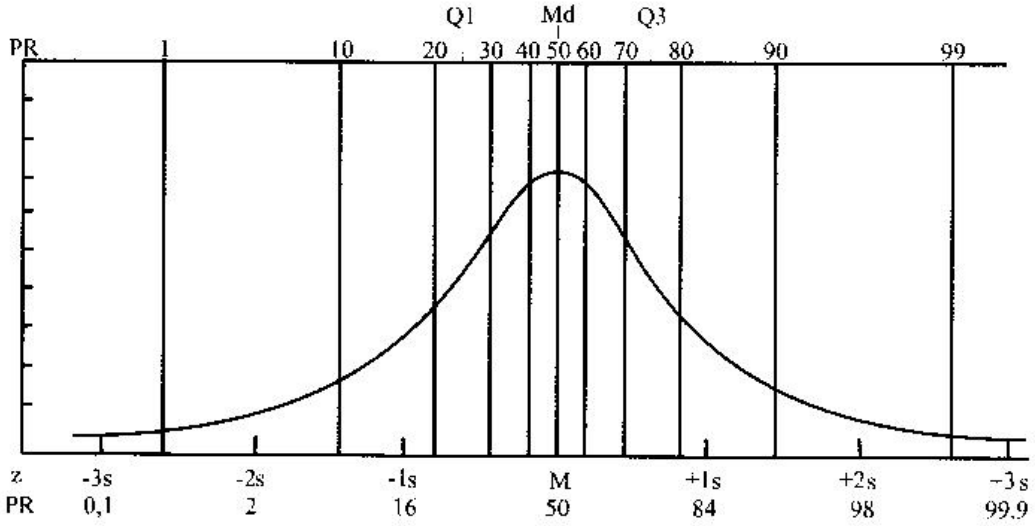


Figura 8-1. Distribuição normal e de postos percentílicos (PR = percentis; Q = quartil; Md – mediana)

O cálculo dos percentis é simples e se encontra ilustrado a seguir na tabela 8-2.

Tabela 8-2. Cálculo dos percentis

Escore bruto (T)	Frequência (f)	Percentagem (%)	Percentagem acumulada
10	4	2	100
9	8	4	98
8	12	6	94
7	32	16	88
6	44	22	72
5	50	25	50
4	30	15	25
3	12	6	10
2	4	2	4
1	2	1	2
0	2	1	1

Para a obtenção dos percentis, basta calcular a percentagem relativa que cada frequência (coluna 2) tem no total de sujeitos, que no caso são 200, e acumular, de baixo para cima, estas percentagens para dar os percentis. Assim, o escore bruto de 6 (coluna 1), que ocorreu 44 vezes em 200 sujeitos (coluna 2) tem percentagem relativa de 22 (coluna 3) e corresponde ao percentil 72 (coluna 4). Quer dizer que 72% dos sujeitos receberam escore menor do que 6. Fica, assim, mais inteligível dizer que o sujeito ganhou escore 72 em vez de 6, porque o 72 é referenciado a uma escala que sempre vai de 0 a 100, enquanto o 6 depende do número de itens do teste.

Os percentis podem ser ilustrados numa ogiva que representa as percentagens acumuladas, onde se lê na abscissa os escores brutos dos sujeitos e na ordenada os seus respectivos escores percentílicos, como na figura 8-2.

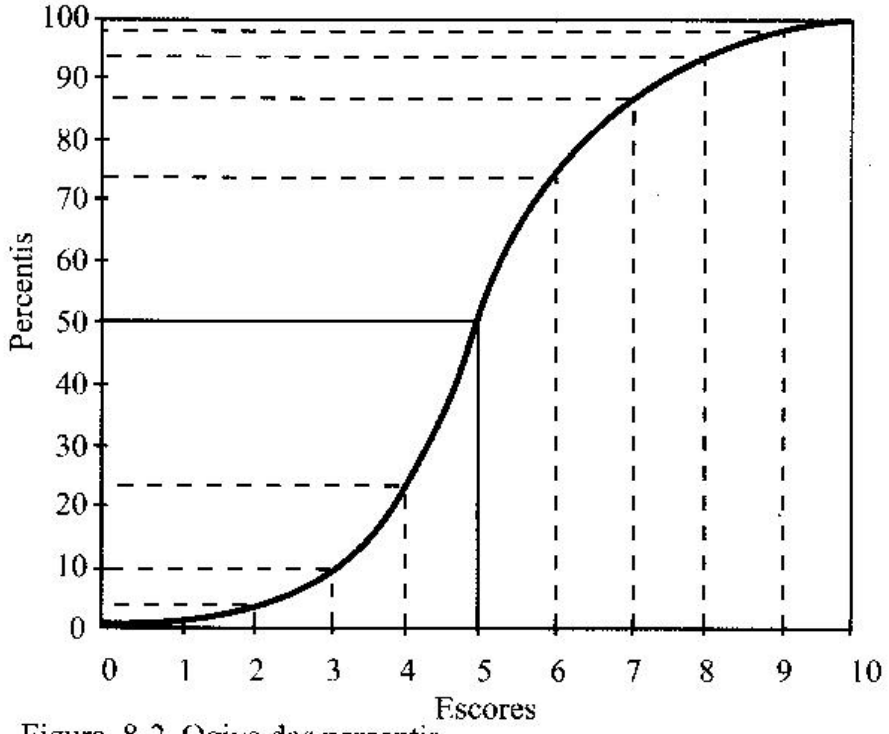


Figura 8-2. Ogiva dos percentis

2.2 – Escore padrão

As normas baseadas no escore padrão (escore *z*) se fundamentam no cálculo deste escore *z* correspondente ao escore bruto do sujeito. Este cálculo é feito de duas formas bastante distintas, que resultarão ou num escore padrão ou num escore padrão normalizado. O primeiro é feito através de uma transformação linear e o segundo através de uma transformação não linear.

2.2.1 – Escore padrão linear é calculado pela fórmula:

$$z = \frac{X - \overline{X}}{s_T} \qquad (8.1)$$

onde,

X = escore bruto do sujeito (o escore T)

$\overline{X}$ = média do grupo no teste

s_T = desvio padrão do teste.

Exemplo. Veja na tabela 8-3 como podemos calcular este escore z.

Tabela 8-3. Cálculo do escore z

Escore bruto (X)	Frequência (f)	fX	Escore padrão (z)
10	2	20	2,55
9	4	36	1,96
8	6	48	1,38
7	16	112	0,79
6	22	142	0,20
5	25	125	-0,39
4	15	60	-0,98
3	6	18	-1,56
2	2	4	-2,15
1	1	1	-2,74
0	1	0	-3,33
Total	100	566	
Média		5,66	
s_T		1,70	

Assim, o escore bruto de 6 terá um escore padrão (z) de 0,20, isto é, ele está 0,20 desvios padrões acima da média, a saber

$$z = \frac{6 - 5,66}{1,70} = 0,20 .$$

2.2.2 – Escore padrão normalizado

Este é calculado através das tabelas da curva normal e consiste, essencialmente, em transformar as percentagens em escores *z*, como ilustrado na figura 8-1 e efetuado na tabela 8-4, novamente com os dados da tabela 8-2.

Tabela 8-4. Cálculo do escore padrão normalizado

Escore bruto (T)	Frequência (f)	Percentagem acumulada	Escore padrão normalizado	Escore padrão (z)
10	2	100	3,29	2,55
9	4	98	2,05	1,96
8	6	94	1,55	1,38
7	16	88	1,18	0,79
6	22	72	0,58	0,20
5	25	50	0,00	-0,39
4	15	25	-0,67	-0,98
3	6	10	-1,28	-1,56
2	2	4	-1,75	-2,15
1	1	2	-2,05	-2,74
0	1	1	-2,33	-3,33

O cálculo do escore padrão normalizado se faz diretamente a partir das tabelas da curva normal, onde, por exemplo, a percentagem acumulada de 98 tem um *z* de 2,05 positivo e a percentagem acumulada de 10 um *z* de -1,28. A utilização desta transformação dos escores brutos é somente justificável se estes se distribuírem normalmente, o que deve ser previamente demonstrado através de algum teste estatístico como, por exemplo, a análise da curtose e da assimetria na distribuição dos escores brutos. Na verdade, se a distribuição dos es-

cores brutos for perfeitamente normal, então os escores padrões normalizados devem ser idênticos aos escores padrões, coisa que, no nosso caso, não é totalmente satisfatório, quando você compara a coluna 4 e 5 da tabela 8-4. Consequentemente, quanto mais as distribuições se afastam da normalidade, menos recomendável é a utilização da transformação não linear dos escores brutos, isto é, não é recomendado transformá-los em escores padrões normalizados.

De qualquer forma que o *z* seja obtido, as normas baseadas nele normalmente utilizam algumas transformações lineares ulteriores para evitar duas dificuldades de uma escala de *z*, a saber: (1) a presença de escores negativos, pois o *z* vai de menos infinito a mais infinito (mais praticamente, de -5 a +5) e (2) a presença de decimais. Para eliminar estas duas deselegâncias, tipicamente o *z* é multiplicado por um coeficiente e ao produto é agregada uma constante, através da fórmula seguinte:

$$T = a + bz$$

onde,

T = escore transformado

z = escore padrão

a, b = constantes quaisquer.

Tanto o coeficiente de multiplicação do *z* (b) quanto a constante somada (a) são arbitrários, resultando em tantas formas de normas derivadas quanto imagináveis. Contudo, alguns dos valores dados a essas constantes são rotineiramente mais utilizados, produzindo normas derivadas que são já tradicionalmente conhecidas, tais como: o escore T, os estaninos, o desvio QI, o escore CEEB (*College Entrance Examination Board*, utilizado para a entrada no ensino superior nos Estados Unidos) e vários outros. As fórmulas de transformação para algumas destas normas são:

$$T = 50 + 10z$$

Desvio QI = 100 + 15z (Escalas de Wechsler) ou

Desvio QI = 100 + 16z (Stanford-Binet)

CEEB = 500 + 100z.

Assim, os escores brutos da tabela 8-3 podem ser expressos em vários tipos de normas derivadas e que são equivalentes, como mostrado na tabela 8-5.

Tabela 8-5. Vários tipos de normas para os mesmos escores brutos

Escore bruto (T)	Escore padrão normalizado	Escore padrão (z)	Escore T	Desvio QI	CEEB
10	3,29	2,55	76	138	755
9	2,05	1,96	70	129	696
8	1,55	1,38	64	121	638
7	1,18	0,79	58	112	579
6	0,58	0,20	52	103	520
5	0,00	-0,39	46	94	461
4	-0,67	-0,98	40	85	402
3	-1,28	-1,56	34	77	344
2	-1,75	-2,15	28	68	285
1	-2,05	-2,74	23	59	226
0	-2,33	-3,33	17	50	167

Todas estas normas são conversíveis umas nas outras, como mostra a figura 8-3. As vantagens de umas sobre as outras é basicamente uma questão de gosto do pesquisador. Apenas é preciso atender ao fato de que as normas percentílicas produzem uma escala ordinal, enquanto os escores padrões e seus derivados dão escalas métricas.

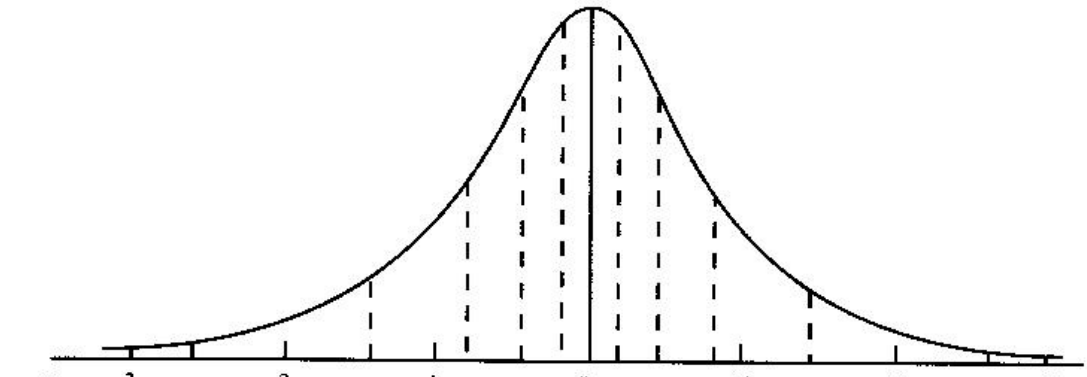


Figura 8-3. Comparação de vários tipos de normas

Popularmente se confundem as escalas percentílicas (P) e as do *z* na hora de se interpretar o escore do sujeito; entretanto, as duas escalas são bastante diferentes. Cf. a figura 8-4 para ver que a escala percentílica é uma transformação não-linear, enquanto a escala T (uma transformação da escala *z*) é uma transformação linear.

Se você compara as duas escalas de normas para o teste de 10 itens da figura 8-4 com uma escala que vai de 0 a 100, você vê que a escala P forma uma ogiva, enquanto a escala T forma uma reta.

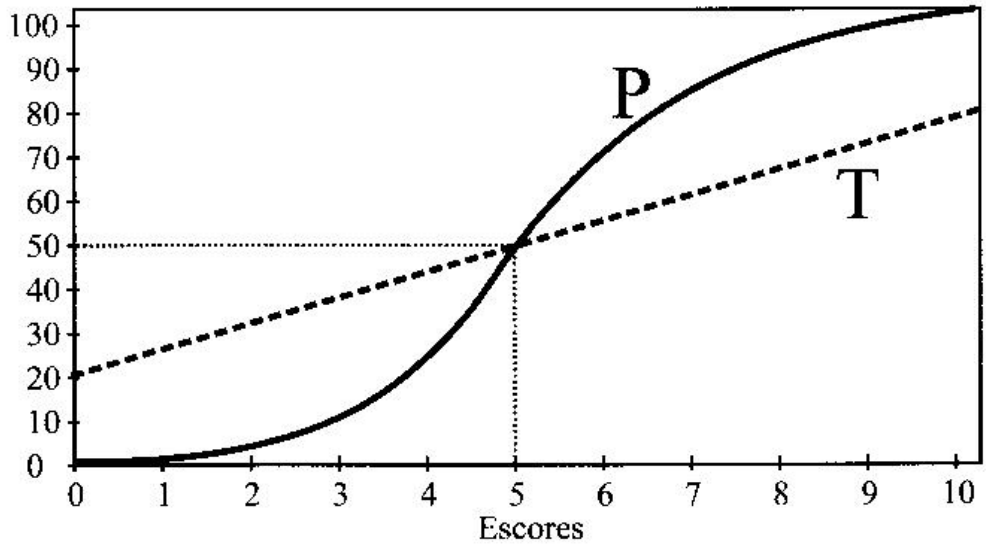


Figura 8-4. Ogiva dos percentis (P) e a reta dos escores T

Mesmo sendo diferentes, os dois tipos de escalas normativas produzem a mesma informação e são conversíveis uma na outra. O problema realmente grave na normatização de um teste não é o tipo de normas utilizado e, sim, a amostra utilizada para fornecer os dados empíricos sobre os quais serão efetuados os cálculos para a produção das normas. Esta amostra, dita grupo normativo, deve ser estatisticamente representativa da população; o que implica, praticamente, em que os grupos normativos são geralmente constituídos de números grandes de sujeitos, implicando em custos proibitivos que tipicamente assustam os pesquisadores ao quererem enveredar em projetos de padronização de testes. Uma rápida inspeção dos testes psicológicos disponíveis no mercado brasileiro dá conta imediata desta precária situação, onde se podem detectar normas baseadas em amostras nunca representativas e de datas quase pré-históricas.

3 – Normas referentes a critério

As normas descritas no ponto anterior são chamadas de normas referentes a grupo, pois com elas quer se comparar o escore de um sujeito com relação a um grupo normativo, isto é, o desempenho do sujeito é comparado com o desempenho do grupo do qual ele faz parte. Tal procedimento é útil quando se desejam discriminar vários níveis de habilidade ou de traços outros numa população. O interesse aqui se dirige para o diagnóstico da variabilidade ou da gama de níveis de habilidade e, por isso, interessa poder diferenciar desde o mais fraco até o mais forte em uma habilidade ou qualquer outro traço latente.

Há, contudo, situações nas quais não interessa ter dados desta natureza, isto é, que discriminem toda uma gama de habilidades ou de personalidade, mas interessa mais decidir se alguém conseguiu ou não um certo nível de habilidade, de aprendizagem ou de traço de personalidade. É tipicamente o caso em provas de domínio de conteúdo, em seleção, em diagnóstico psiquiátrico e também em orientação vocacional.

Assim, por exemplo, em situação de seleção de recursos humanos ou em avaliação de treinamento, não interessa saber os escores individuais que todos os candidatos pudessem obter, mas interessa saber se os candidatos atingirem um critério preestabelecido de desempenho. Exemplificando: num teste de conhecimento ou de inteligência, interessa somente saber quem obteve escore X, que corresponde, vamos dizer, ao escore que somente 20% da população seria capaz de obter ou um escore que corresponde ao desempenho de quem foi capaz de dominar 80% do conteúdo de um treinamento. Assim, a norma nestes casos seria a obtenção de 80% do conteúdo de um treinamento, porque somente tal percentual representaria domínio do conteúdo, não interessando o percentual de domínio de conteúdo que cai abaixo dos 80%.

Similarmente, em testes psiquiátricos, normalmente o que interessa é saber se um dado sujeito atingiu o escore que o põe na faixa de sujeitos psiquiátricos *vis-à-vis* os normais. Novamente, não interessa saber o perfil de personalidade do sujeito, mas sim se ele atingiu ou não o ponto-limítrofe de normalidade-anormalidade. É o caso de testes como MMPI que tem para cada escala um ponto crítico que indica a linha divisória de passagem para a anormalidade.

Neste contexto, são muito utilizadas normas do tipo acima exposto. Dois destes tipos são particularmente úteis e utilizados, a saber, no caso dos testes referentes a critério (especificamente em educação) e as normas via tabelas de expectância. O primeiro tipo define as normas teoricamente e o segundo através de dados empiricamente verificados.

3.1 – Testes referentes a critério

Esta tendência surgiu com Glaser (1963), que pela primeira vez utilizou a expressão *criterion-referenced testing* para designar o tipo de testes, particularmente no campo da educação, que vinham sendo utilizados sob outras expressões, tais como testes referentes a conteúdo, testes referentes a domínio e testes referentes a objetivos. Todas estas expressões sinalizam o critério que era utilizado na interpretação dos escores de uma prova educacional. Isto é, as provas ou testes eram criados para medir um certo conteúdo, ou um certo conjunto de objetivos educacionais (os processos cognitivos, do tipo estabelecido pelas taxonomias, tais como as de Bloom – Bloom, 1956; Bloom, Hastings & Madaus, 1971 – e de outros) ou, mesmo, para verificar se um conteúdo instrucional específico fora dominado pelo aluno, a saber, se ele adquirira o domínio completo do dado conteúdo programático. Neste caso, domínio completo significa um domínio de cerca de 80% ou mais do conteúdo especificado pelo programa.

Nestes casos, o que está em jogo é, primeiramente, a explicitação detalhada do conteúdo programático em tópicos e subtópicos, bem como da percentagem em que cada um deles deve estar representado, em termos de importância, no teste (testes referentes a conteúdo) ou a explicitação dos objetivos ou processos cognitivos envolvidos na aprendizagem de tal conteúdo (testes referentes a objetivos), isto é, elaborar as tabelas de especificação e, em seguida, construir os itens para cobrir adequadamente os tópicos ou objetivos assim explicitados (cf. capítulo 6 sobre validade de conteúdo).

O fundamento que alicerça o conceito do domínio de aprendizagem é de que os investimentos em educação, por exemplo, devem resultar num aprendizado que vai além do ensaio e erro. Espera-se, na verdade, que com a intervenção da educação, no final os alunos dominem o conteúdo e não apenas que o resultado final mostre que o nível de aprendizado se distribua dentro da curva normal, onde joga a lei da aleatoriedade, isto

é, a maior parte dos alunos aprende medianamente, alguns não aprendem nada e outros aprendem tudo. O objetivo da educação é de que, no final, *todos* os alunos aprendam tudo, isto é, dominem o conteúdo. Então, obviamente os resultados finais são todos enviesados para a cauda direita da curva normal, isto é, a distribuição não pode ser normal; aliás, se ela sai normal é indicação suficiente de que não houve aprendizagem. Como diz Carroll (1963), se os alunos se distribuem normalmente no que respeita à distribuição das aptidões, os resultados do ensino que se adapta às habilidades individuais de cada aluno devem apresentar-se iguais para todos os alunos de todos os níveis de habilidades, isto é, todos os alunos no final terão dominado o conteúdo e, por consequência, os escores serão todos os mesmos, resultando numa distribuição de escores de tipo *J* e não da curva normal (cf. figura 8-5).

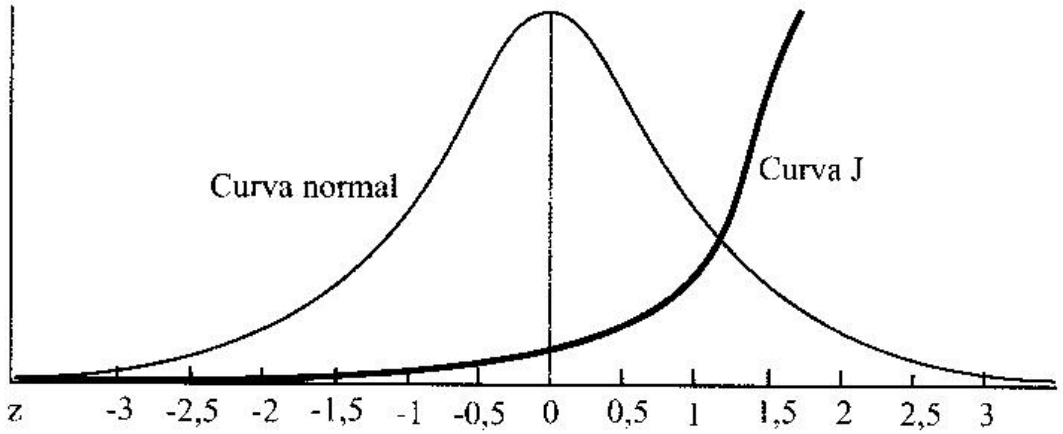


Figura 8-5. Comparação da curva normal com a curva J

Em casos como este, as normas a serem utilizadas para a interpretação dos escores (resultados) não podem ser as da curva normal nem as percentílicas, mas sim normas referentes a critério, sendo este critério definido teoricamente, preliminarmente ao próprio conhecimento dos resultados efetivos na prova ou teste. Este critério, como dissemos, é tipicamente colocado em torno de 80% de domínio do conteúdo (programático e de objetivos).

Raciocínio similar se usa no caso dos testes psiquiátricos, através dos quais o interesse se situa em definir quem atingiu e quem não atingiu um critério X que determina o limiar entre normalidade e anormalidade. Cada teste deverá definir o escore que define este limiar. Por exemplo, no

caso do Questionário de Saúde Geral de Goldberg (Goldberg, 1972), este critério é o valor 3 na escala que vai de 1 a 4.

3.2 – Tabelas de expectância

Diferentemente das normas discutidas sob testes referentes a critério, onde elas são estabelecidas teoricamente, *a priori*, no caso das tabelas ou gráficos de expectância as normas são determinadas diretamente por dados empíricos. Estas tabelas dão a probabilidade de êxito dos sujeitos num dado campo de atividade em função do escore que eles obtêm num dado teste. Veja o exemplo dado por Flanagan (1947) na figura 8-6.

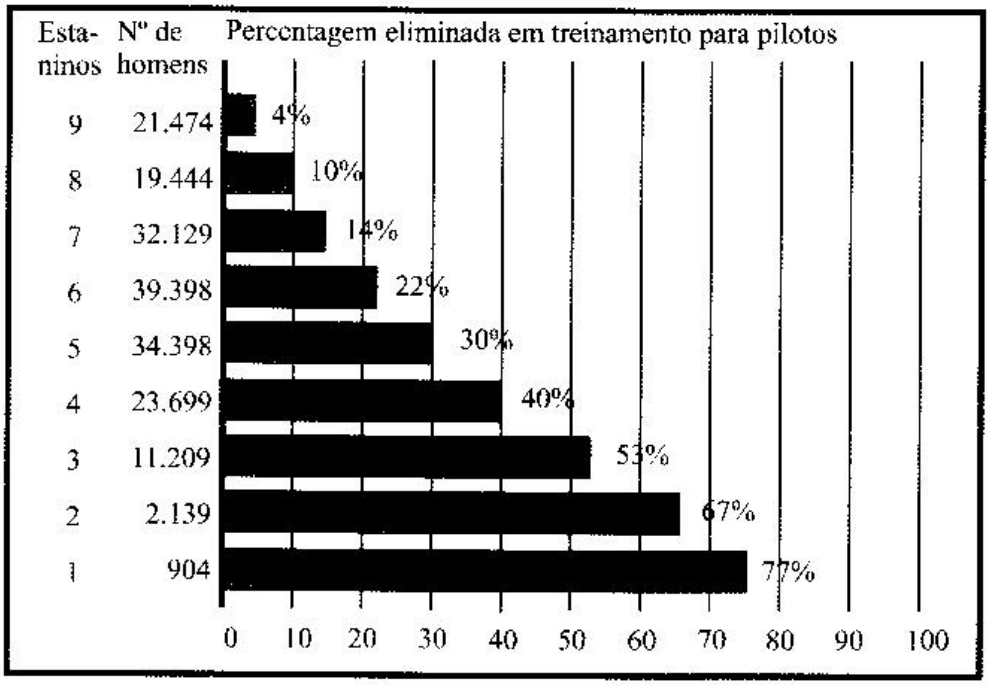


Figura 8-6. Tabela de expectância mostrando a relação entre desempenho em uma bateria de testes de seleção para pilotos e a eliminação do treinamento de voo

Esta figura mostra os resultados, em estaninos, obtidos por cadetes da aeronáutica norte-americana e a correspondente percentagem daqueles que foram eliminados do treinamento de voo. Como se vê, há uma relação direta entre baixos escores estaninos e alta percentagem de eliminação. Assim, dos cadetes que obtiveram estanino 5, 30% deles foram eliminados, ficando 70% para o treinamento de voo, ao passo que aqueles que obtiveram um estanino de 9, apenas 4% deles foram eliminados do treinamento. Desta forma, com um tal gráfico de dados, pode-se predizer o percentual de êxito numa situação, conhecendo-se o escore que o sujeito obteve num dado teste. De fato, por exemplo o sujeito que obteve,

neste caso dos pilotos, um estanino de 4 tem uma chance de 60:40 ou de 3:2 de completar o curso de treinamento. É preciso, contudo, lembrar que tabelas desta natureza são elaboradas em cima de resultados obtidos empiricamente e não decididos a priori e, por isso, é importante dar total atenção ao parágrafo seguinte.

Tais tabelas são relativamente fáceis de construir; contudo, para elaborar tabelas de expectância válidas, deve-se trabalhar com amostras grandes e representativas da população, do contrário tais tabelas se tornam totalmente inadequadas e graves injustiças se podem cometer numa seleção com base em tais dados. Em posse de tabelas adequadas, entretanto, é largamente facilitado o trabalho de decisão sobre a interpretação de um escore no teste. Por exemplo, podemos construir a seguinte tabela de expectância (tabela 8-6) entre resultados no teste Raven e as menções escolares (dados fictícios).

Tabela 8-6. Exemplo de tabela de expectância

Raven	Percentual recebendo as menções escolares				
	II	MI	MM	MS	SS
00-10	50	0	10	1	0
10-20	20	30	10	2	0
20-30	15	20	20	10	2
30-40	10	10	30	17	15
40-50	5	7	20	30	33
50-60	0	3	10	40	50

Com base nesta tabela 8-6, podemos saber que sujeitos que obtiveram um escore entre 40 e 50 no Raven têm uma chance de 33% de obter uma menção escolar de SS, 30% MS, 20% MM, 7% MI e 5% II. Ou, o sujeito que obteve máximos no Raven (50-60) terá chance de 50% de obter uma menção de SS.

4 – Normatização na TRI

A TRI dá o nível de teta para cada sujeito. O valor θ de um sujeito define a probabilidade dele acertar um dado item. A escala θ, onde se situa o escore do sujeito, vem expressa em termos de escores padrões com média = 0 e desvio padrão = 1 e uma amplitude que vai de -∞ a +∞, ou mais praticamente, de -3 a +3. Essa escala é de difícil intelecção para a maioria do público, além de ser deselegante. Assim, para facilitar o uso prático, a escala do θ pode ser transformada em escalas mais apropriadas e mais inteligíveis, fazendo uso de transformações lineares, não lineares ou transformá-la em escores T que são de uso mais corrente e, consequentemente, de mais fácil compreensão. Vejamos um pouco essa estória.

As *transformações lineares* consistem em expressar o θ acrescentando-se a ele uma constante *k* e, ainda, além de uma constante também um fator multiplicativo *m*. Contudo, ao se fazer tal operação no θ, é necessário que se faça o mesmo nos parâmetros dos itens para, assim, manter a invariância da escala θ.

Temos, assim, que

Para o modelo de 1 parâmetro: $\theta' = m(\theta) + k$

$b' = m(b) + k$

Para o modelo de 2 parâmetros: $\theta' = m(\theta) + k$

$b' = m(b) + k$

$a' = a/m$

Para o modelo de 3 parâmetros: $\theta' = m(\theta) + k$

$b' = m(b) + k$

$a' = a/m$

$c' = c.$

Por exemplo, Woodcock (1978) fez a seguinte transformação para seu teste: $w_\theta = 20 \log_9(e^\theta) + 500$

$= 9{,}1\theta + 500$

e

$w_b = 9{,}1b + 500.$

Essa é uma transformação feita para uma escala logarítmica de base 9. Veja, então, como surgiu essa estranha transformação de Woodcock:

$$\log_9(e^{\theta}) = \theta \log_9 e = 0{,}455$$

$$20 \log_9(e^{\theta}) = 20 \times 0{,}455 = 9{,}1.$$

Desta forma, 9,1 é o fator multiplicativo da transformação de Woodcock e a constante acrescida é 500, resultando numa escala normatizada que tem como a média 500, o que torna essa escala parecida com a escala CEEB, a qual, como vimos anteriormente, é uma escala resultante de uma transformação da escala *z*.

Wright (1977) modificou a escala de Woodcock e a chamou de WITs. Essa modificação simplesmente quis tornar a nova escala parecida com as escalas dos índices QI, as quais têm uma média de 100. Cf. a fórmula:

$$\text{WITs} = 9{,}1\theta + 100.$$

Quanto às *transformações não lineares*, deve-se dizer que há uma série delas; contudo, normalmente elas não facilitam a interpretação dos escores, o que, afinal, é a razão principal das transformações. De qualquer forma, a mais utilizada é a transformação *logits*, que faz uso de logaritmos naturais de base *e*. No caso do modelo de 1 parâmetro, ela seria

$$\ln \frac{P(\theta)}{Q(\theta)} = \theta - b$$

onde $Q(\theta) = 1 - P(\theta)$.

Entretanto, a transformação mais interessante é a *transformação do θ no escore verdadeiro (τ)* da Psicometria Clássica.

Você se lembra que o escore verdadeiro V na TCT foi definido como a esperança matemática do escore bruto T (cf. cap. 4), sendo este a soma dos acertos dos itens de um teste, ou seja,

$$T = \sum_{i=1}^{n} u_i$$

onde, u_i é a resposta de acerto (= 1) ou erro (= 0) dada ao item *i*.

Agora, a esperança matemática do escore T é V ou τ, ou seja,

$$\tau = V = E(T) = E\left(\sum_{i=1}^{n} u_i\right), \text{ ou seja (cf. encarte),}$$

$$\tau = \sum_{i=1}^{n} P_i(\theta).$$

Esta fórmula surge do seguinte:

$$\tau = E(T) = E\left(\sum_{i=1}^{n} u_i\right)$$

$$E\left(\sum_{i=1}^{n} u_i\right) = \sum_{i=1}^{n} E(u_i)$$

Mas, se uma variável, como o caso do u_i, só assume dois valores u_1 e u_2, sendo suas probabilidades de P_1 e P_2, então

$$E(u) = P_i u_1 + P_2 u_2$$

Como, em testes de aptidão, as respostas somente valem 1 (acerto) ou 0 (erro), que são expressas como $P_i(\theta)$ com probabilidade 1 e $Q_i(\theta) = 1 - P_i(\theta)$, então

$$E(u_i) = 1 \times P_i(\theta) + 0 \times Q_i(\theta) = P_i(\theta)$$

Assim, $\tau = \sum_{i=1}^{n} P_i(\theta)$.

Assim, o escore verdadeiro τ e o q estão monotonicamente relacionados, mas a transformação é não linear, pois enquanto $P_i(\theta)$ vai de 0 a 1 o τ vai de 0 a n (número de respostas corretas).

A transformação da escala θ na escala τ dá o escore T, o qual pode ser transformado numa escala percentílica, que é sempre mais fácil de entender. Para tanto basta dividir este escore verdadeiro pelo número de itens e aí surge a escala percentílica. Cf. a fórmula:

$$\pi = \frac{1}{n} \sum_{i=1}^{n} P_i(\theta).$$

Esta transformação muda o escore τ em proporções, que multiplicadas por 100 dão a escala percentílica.

Obs.: note as letras gregas τ e π nestas fórmulas; elas são para alertar que o escore verdadeiro e a escala percentílica são transformações derivadas do θ.

Um exemplo (adaptado de Hambleton, Swaminathan & Rogers, 1991):

Considere um teste de 5 itens com os parâmetros da tabela 8-7 respondidos por sujeitos cujos tetas variam de -3 a +3.

Tabela 8-7. Parâmetros de 5 itens

Item	a_i	b_i	c_i
1	0,80	-2,00	0,00
2	1,00	-1,00	0,00
3	1,20	0,00	0,10
4	1,50	1,00	0,15
5	2,00	2,00	0,20

Calculando as probabilidades de acerto dos 5 itens por sujeitos de habilidades diferentes, dá os resultados da tabela 8-8, onde também é feita a transformação dos resultados em τ (tau) e π (proporção) e P (percentil).

Tabela 8-8. Relação entre θ e π

Para θ	Probabilidades dos 5 itens					$\tau =$	$\pi =$	P =
(aptidão)	$P_1(\theta)$	$P_2(\theta)$	$P_3(\theta)$	$P_4(\theta)$	$P_5(\theta)$	$\Sigma P_i(\theta)$	τ/n	100π
-3	0,20	0,03	0,10	0,15	0,20	0,69	0,14	14
-2	0,50	0,15	0,11	0,15	0,20	1,12	0,22	22
-1	0,80	0,50	0,20	0,16	0,20	1,85	0,37	37
0	0,94	0,85	0,55	0,21	0,20	2,75	0,55	55
1	0,98	0,97	0,90	0,58	0,21	3,65	0,73	73
2	0,99	0,99	0,99	0,94	0,60	4,51	0,90	90
3	1,00	1,00	1,00	1,00	0,96	4,96	0,99	99

5 – Expressão dos escores em faixas

Existe na literatura uma louvável tendência de se apresentarem os resultados dos sujeitos num teste em termos de faixas definidas pelo erro padrão de medida (EPM) em vez de escores isolados. Esta prática permite, igualmente, comparar com maior precisão a diferença entre dois escores, observando se as faixas dos dois escores se sobrepõem ou não.

Um procedimento de medida qualquer, por exemplo o escore em um teste, expressa uma variabilidade de fatores, parte 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. Vimos no capítulo sobre a fidedignidade dos testes que esta 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* à 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, que designa o erro provável de medida incorrido pelo teste. Este erro, como você se lembra, é definido em termos padrões e é o seguinte: $EPM = s_T\sqrt{1-r_{tt}}$ onde, o erro padrão da medida (EPM) se expressa em termos do desvio padrão do teste (s_T) e do coeficiente de precisão do mesmo teste (r_{tt}) obtidos na mesma amostra de sujeitos. Veja o exemplo da tabela 8-7. Neste exemplo, o $s_T = 1,7$ e a fidedignidade $r_{tt} = 0,94$. Desta forma, o $EPM = 1,7\sqrt{1-0,94} = 0,42$. Assim, a faixa de pontuação do sujeito vai ser seu escore bruto +0,42 e -0,42. Os resultados estão na tabela 8-9.

Tabela 8-9. Escores brutos expressos em faixas

Escore bruto	Frequência	Faixa
10	2	9,58 – 10,00
9	4	8,58 – 9,42
8	6	7,58 – 8,42
7	16	6,58 – 7,42
6	22	5,58 – 6,42
5	25	4,58 – 5,42
4	15	3,58 – 4,42
3	6	2,58 – 3,42
2	2	1,58 – 2,42
1	1	0,58 – 1,42
0	1	0,00 – 0,42
N	100	
Média	5,66	
s_T	1,70	

De fato, fazendo a suposição de que os erros de medida se distribuem normalmente, este índice se apresenta muito útil na interpretação de escores empíricos individuais dos sujeitos, pois com ele se pode definir os limites do intervalo dentro do qual mais provavelmente se situa o escore verdadeiro do sujeito.

CAPÍTULO 9
Equiparação de escores

Introdução

A equiparação ou equivalência dos escores (*equating* em inglês) tem como tema resolver o problema de se compararem escores obtidos pelos mesmos sujeitos em testes diferentes que, obviamente, estejam medindo o mesmo construto, ou, se quiser, escores obtidos em mais de uma forma de um mesmo teste. Equiparação de escores é um processo estatístico que serve para ajustar escores em formas diferentes de um teste de tal sorte que os escores em ambos os testes possam ser utilizados indiferentemente (Kolen & Brennan, 1995). Angoff (1984) simplifica, dizendo que equalização serve para converter um sistema de unidades da forma de medida de um teste para o sistema de unidades de uma outra forma do mesmo teste, assim como existe um método de conversão do sistema de unidades de medida da temperatura Celsius para o sistema de unidades de medida Fahrenheit. Ela visa ajustar níveis diferentes de dificuldade entre as formas do teste, que foram criadas para serem similares tanto em dificuldade quanto em conteúdo, mas que na prática elas não têm o mesmo nível de dificuldade. Por exemplo, um sujeito recebeu dois escores em raciocínio verbal, um no teste A, obtendo o escore de 50, e um no teste B, obtendo 55. A pergunta que se levanta é se estes dois escores podem ser diretamente comparados, isto é, os dois escores estão refletindo o mesmo nível de aptidão do sujeito ou os dois escores podem ser utilizados indiferentemente por estarem a dar a mesma informação sobre o sujeito? E como deve ser expresso o fato dos dois escores serem equivalentes? Ou, como expressar escores assim diferentes numa mesma escala?

Esta situação é muito comum em situações de avaliação de desempenho, onde testes diferentes, ou formas diferentes de um mesmo teste, são aplicados aos mesmos sujeitos para estimar seu aproveitamento. Como os testes são diferentes cada vez que se avalia o desempenho dos

mesmos sujeitos e podendo produzir escores diferentes para os mesmos sujeitos ou, mesmo, sujeitos diferentes tomando formas diferentes de um mesmo teste, fica-se perguntando como se pode comparar tais escores.

Com o uso maciço dos testes em situações acadêmicas e de seleção, este problema tem ultimamente recebido muita atenção (Lord, 1980; Holland & Rubin, 1982; Angoff, 1984; Hambleton & Swaminathan, 1985; *Applied Psychological Measurement*, 1987; Kolen, 1988; Hambleton, Swaminathan, & Rogers, 1991; Muñiz, 1992; Kolen & Brennan, 1995; Rabello, 2001). As respostas dadas ao problema são variadas e algumas muito complexas, como veremos. Aliás, inclusive a nomenclatura já começa a ser abundante nesta área; você vai encontrar expressões tais como: ancoragem, *bench marking*, equalização por calibração, equalização horizontal, equalização vertical, predição, escalagem, moderação estatística, propósitos sociais (Suathong, Schumacker & Beyerlein, 2000; Mislevy, 1992; Linn, 1993). Em inglês também você tem várias expressões, tais como: *equating, linking, scaling*. Embora essas expressões indiquem propósitos diferentes, elas todas se referem a técnicas cujo objetivo fundamental consiste em se justificar e operacionalizar funções de ligação (*linking*) entre dois eventos diferentes, mas relacionados, tais como os escores em dois testes diferentes que medem a mesma coisa.

Para resolver tais problemas de diferenças entre escores obtidos em formas diferentes de um teste, utiliza-se a equiparação de escores, a qual é um processo estatístico e que requer certos tipos de delineamento na coleta da informação para permitir o uso dos vários métodos estatísticos de *equating* existentes. Há basicamente três modelos de delineamentos de coleta de dados para fazer a equiparação dos escores dos sujeitos em diferentes testes. Os delineamentos são:

- um grupo único contrabalançado: um grupo randômico e duas formas do teste;
- grupos equivalentes (randômicos): dois grupos randômicos e um teste;
- teste de ancoragem (grupos não equivalentes com itens comuns): dois grupos, dois testes e um teste de ancoragem.

1 – Os delineamentos

Os delineamentos dizem respeito a como são coletados os dados que servirão para o estudo da equiparação de escores: são eles coletados de uma mesma amostra dos mesmos sujeitos ou são utilizadas amostras diferentes de sujeitos?

1.1 – Um grupo único contrabalançado

É o caso no qual dois ou mais testes são aplicados a uma amostra aleatória dos mesmos sujeitos. Os testes podem ser aplicados na mesma ocasião ou, normalmente, um aplicado numa ocasião e o outro teste numa ocasião posterior. Entretanto, para evitar efeitos de sequência das formas do teste, normalmente são formados dois subgrupos de sujeitos, a um deles é dado o caderno que contém a forma A seguida pela forma B, sendo o inverso para o subgrupo 2. Para efetuar este delineamento, tipicamente se usa o procedimento em espiral, que é o seguinte:

- dois cadernos de testes são construídos (X e Y); um caderno (X) contendo a Forma A seguida da Forma B do mesmo teste e outro (Y) contendo a Forma B seguida da Forma A (cf. figura 9-1);
- seleciona-se um grupo aleatório de sujeitos;
- distribuem-se os cadernos da seguinte maneira: ao primeiro sujeito é dado o caderno X, ao segundo o caderno Y, ao terceiro o caderno X e assim por diante.

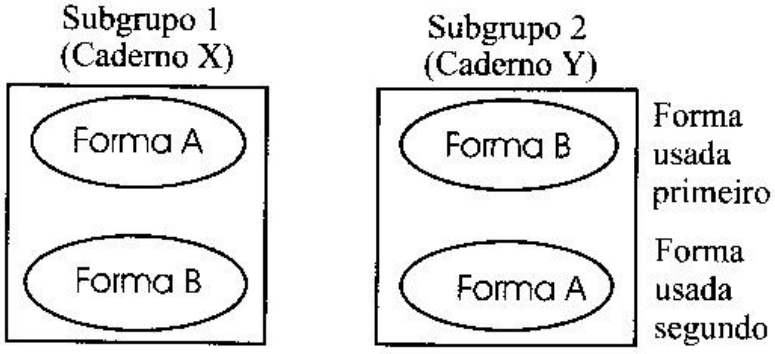


Figura 9-1. Grupo único contrabalançado

Desta maneira se controla a ordem de apresentação das formas do teste, sendo as duas formas do teste, A e B, equivalentes quanto a este aspecto, o que não se poderia assumir se os mesmos sujeitos tivessem sido

expostos à Forma A primeiro e mais tarde à Forma B. Neste último caso, se aparecerem diferenças nos escores entre as duas Formas, elas poderiam ser simplesmente um efeito de ordem de apresentação. Mas no caso do formato em espiral, pode-se verificar os efeitos de ordem de apresentação, como também diferenças reais nos escores entre as duas formas. Cf. o exemplo da tabela 9-1.

Tabela 9-1. Médias em duas formas de um teste para um grupo único contrabalançado

	Subgrupo 1	Subgrupo 2
Ordem de apresentação das formas	*Forma A*: 64	*Forma B*: 65
	Forma B: 67	*Forma A*: 62

Pelos dados da tabela 9-1, as médias para a forma do teste aplicada em primeiro lugar foram de 64 para o subgrupo 1 na Forma A e de 65 para o subgrupo 2 na Forma B; uma diferença de 1 ponto em favor da Forma B. Esta diferença sobe para 5 quando as formas são aplicadas em segundo lugar. Em casos como este, Kolen e Brennan (1995) aconselham que as formas aplicadas em segundo lugar sejam simplesmente descartadas, mantendo-se as aplicadas em primeiro lugar para as análises de equiparação.

Para se poder trabalhar com este delineamento, o número de sujeitos deve ser razoavelmente grande, objetivando dar conta da variância erro que ocorre na estimação da equiparação dos escores. Kolen e Brennan (1995: cap. 7) dão fórmulas complicadas para decidir quão grande deve ser a amostra. Em geral, uma amostra de 500 será adequada.

Precaução: Para que o procedimento em espiral funcione, isto é, que produza grupos aleatórios, os respondentes não podem ter sido distribuídos na sala de uma forma sistemática, como por exemplo menino – menina, menino – menina, etc.

1.2 – Grupos randômicos (equivalentes)

Neste delineamento são formados dois grupos aleatórios; um destes grupos responde à Forma A e o outro à Forma B. Para formar estes grupos, pode-se utilizar o procedimento em espiral: o primeiro sujeito recebe a Forma A, o segundo a Forma B, o terceiro a Forma A, etc. Uma vantagem deste delineamento é que cada grupo responde a somente uma forma do teste, reduzindo o tempo e o cansaço da testagem. Inclusive, pode-se utilizar mais de duas formas do teste neste processo de espiral. Veja na figura 9-2 o formato deste delineamento.

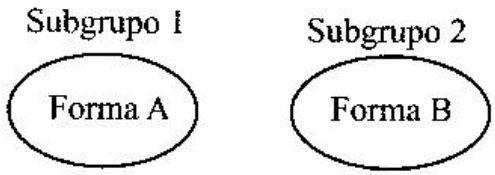


Figura 9-2. Grupos randômicos

1.3 – Grupos não equivalentes com teste de ancoragem

Este é o delineamento mais utilizado (cf. figura 9-3) e, talvez, mais apropriado para fazer a equiparação dos escores. Consiste no seguinte: dois testes, A e B, são aplicados a duas amostras diferentes de sujeitos na mesma ocasião, como no caso anterior, apenas que em ambos os testes A e B existe um número X de itens que são idênticos (os mesmos itens em ambos os testes). Assim, de fato temos três testes: o A, o B e o X, este último incluído tanto no A quanto no B. Note porque, neste caso, as duas amostras de sujeitos não precisam ser equivalentes, pois temos o teste X, que é idêntico para ambas as amostras, que precisamente servirá de referência na equiparação dos escores.

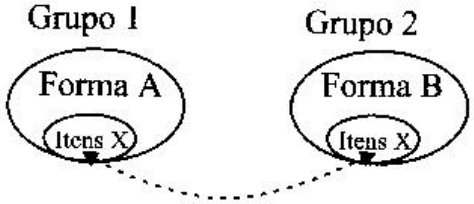


Figura 9-3. Teste de ancoragem

O conteúdo do teste de ancoragem X é algo muito crítico neste delineamento, porque dele vai depender a possibilidade de se poder separar os efeitos diferenciais na aptidão (dos dois grupos) das diferenças em dificuldade dos dois testes A e B. Desta forma, o conteúdo do teste de ancoragem deve ser representativo tanto do teste A quanto do teste B e ser incluído exatamente idêntico em ambos os testes. Sendo assim, então a diferença observada entre os escores dos dois grupos no teste de ancoragem X irá dizer qual dos dois grupos tem maior aptidão, dado que deve ser levado em conta na hora que se vai fazer a equiparação dos dois testes A e B. Isto é, se o grupo 2, por exemplo, é em média 10% mais hábil que o grupo 1 no teste de ancoragem X, então estes 10% de maior aptidão do grupo 2 deve ser descontado quando se fizer a equiparação entre as dificuldades relativas dos dois testes A e B.

2 – Os métodos

Para fazer a equiparação entre os escores de duas formas de um teste ou entre dois testes paralelos, pode-se transformar a escala de uma das formas na outra através de equações, das quais existem três categorias básicas, a saber:

– Equações lineares (equiparação linear – *linear equating*)

– Equações não lineares (equiparação equipercentílica – *equipercentile equating*)

– Equiparação dentro da TRI (Teoria de Resposta ao Item).

Dentro de cada uma dessas categorias existem vários métodos de equiparação, sobre alguns dos quais falaremos a seguir.

2.1 – Métodos lineares

Na equiparação linear de escores (*linear equating*) se postula que o escore x no teste X e o escore y no teste Y estão linearmente relacionados, isto é,

$$y = ax + b \qquad (9.1)$$

Esta fórmula representa uma linha de regressão entre y e x; nela o a representa a inclinação da curva (*slope*) e o b o intercepto da mesma, que são os dois parâmetros que definem uma linha de regressão. Para

descobrir o *a* e o *b* dessa fórmula, você deve levar em conta as médias (M_X e M_Y) e os desvios-padrão (s_Y e s_X) de ambas as distribuições (X e Y), que também estão relacionados linearmente, ou seja,

$$M_Y = aM_X + b$$

$$s_Y = as_X$$

Disso segue que

$$a = \frac{s_Y}{s_X} \quad \text{e}$$

$$b = M_Y - \frac{s_Y}{s_X} M_X$$

Desta forma, a equação 9.1 se torna a seguinte:

$$y = \frac{s_Y}{s_X}(x - M_X) + M_Y \qquad (9.2)$$

Em vez de trabalhar com escores brutos, você pode utilizar os escores padrão dos dois testes, isto é (cf. ponto 2.1.2 mais adiante),

$$z_Y = z_X \quad \text{ou seja,} \quad \frac{x - M_X}{s_X} = \frac{y - M_Y}{s_Y}$$

Para o caso da transformação linear, você pode considerar estas duas possibilidades (cf. Angoff, 1984; Woldbeck, 1998):

1) Se as diferenças entre as duas formas do teste são consideradas iguais ao longo de toda a escala, isto é, a dificuldade de uma forma é igualmente maior tanto para escores baixos quanto médios e altos, então faça a *equiparação pelas médias*. Normalmente, esta suposição é difícil de ser sustentada;

2) Se, contudo, você considerar que as diferenças entre as duas escalas não são idênticas ao longo de toda a extensão da escala, então você pode fazer uma equiparação linear, transformando os escores brutos em escores *z*.

2.1.1 – Equiparação pela média

Vamos considerar duas formas, Forma A e Forma B, de um mesmo teste. Se a Forma A difere da Forma B em dificuldade de maneira igual ao longo de toda a escala, para equiparar os escores basta acrescentar uma constante à escala da forma mais difícil. Parte-se da ideia de que escores, nas duas formas, que se situam a distâncias iguais das suas respectivas médias podem ser igualados. Assim:

$$A - M_A = B - M_B \tag{9.3}$$

Que resolvendo para B vai dar

$$B = A - M_A + M_B \tag{9.4}$$

Onde,

A e B são os escores na Forma A e na Forma B respectivamente

M_A e M_B são as médias dos escores nas duas formas.

Por exemplo: Se a média da Forma A foi de 70 e a da Forma B de 60, então temos que:

$B = A - 70 + 60 = A - 10.$

Desta forma, um escore de 60 na Forma B é considerado indicar o mesmo nível que um escore de 70 na Forma A (60 = 70 – 10). De sorte que, para equiparar as duas formas do teste, você deverá ou acrescentar 10 pontos ao escore obtido na Forma B ou subtrair 10 pontos da Forma A.

Este procedimento somente faz sentido se os dois testes tiverem a mesma distribuição em termos de curtose e de assimetria, isto é, as duas distribuições ou são ambas normais, diferindo apenas em termos de dificuldade (a média de um é maior que a do outro teste) ou ambas são igualmente leptocúrticas ou platicúrticas, além de apresentarem os mesmos índices de assimetria. Hoje em dia, com tantas outras técnicas mais pertinentes para proceder à equiparação dos escores, esta técnica da equiparação pela média já não se justifica mais e é, de fato, raramente utilizada.

2.1.2 – Equiparação linear

Se a diferença dos escores entre as duas formas do teste for diferente em diferentes pontos da escala, então uma transformação linear entre as duas formas é recomendada. Neste caso, procura-se igualar não as médias, mas sim os escores-padrão das duas distribuições. Esta técnica consiste em determinar qual escore z de um teste corresponde ao escore z do outro teste, levando em conta as distribuições específicas de cada teste. Assim, o que se deseja é verificar a equiparação $z_A = z_B$, isto é, é preciso verificar qual é o escore z da Forma B que corresponde, é equivalente, ao z da Forma A. Então temos,

$$z_A = z_B, \text{ ou seja, } \frac{A - M_A}{s_A} = \frac{B - M_B}{s_B} \qquad (9.5)$$

que, resolvendo para B vai dar

$$B = s_B\left(\frac{A - M_A}{s_A}\right) + M_B \text{ ou, rearranjando os termos, temos} \qquad (9.6)$$

$$B = \frac{s_B}{s_A}A + \left[M_B - \frac{s_B}{s_A}M_A\right]$$

a qual é a equação linear de conversão dos escores da Forma B para a escala da Forma A. Onde, s_A e s_B são os desvios-padrão (DP) das duas formas do teste; o resto como na fórmula 9.4.

O rearranjo na fórmula 9.6 é para mostrar que a fórmula expressa a inclinação (primeira parte da fórmula) e o intercepto (segunda parte da fórmula) da linha de regressão da Forma B sobre a Forma A, ou seja,

– o $\frac{s_B}{s_A}A$ é a *inclinação* da reta de regressão da Forma B em A e

– $M_B - \frac{s_B}{s_A}M_A$ é o *intercepto* da reta.

Desta forma, a equação a ser resolvida será:

B = inclinação de A + intercepto.

Para resolver esta equação temos que ter em mãos os seguintes cinco dados:

– M_A : a média do teste A

– M_B : a média do teste B

– s_A : o desvio-padrão de A

– s_B : o desvio-padrão de B

– um escore A: escore de um sujeito no teste A.

Para esta situação, Angoff (1984) desenvolveu uma fórmula para o cálculo do erro padrão, que é a seguinte:

$$s_E = \sqrt{\frac{2s_B^2}{N_t}(z_A + 2)}$$

onde, N_t é o número total de sujeitos de ambas as amostras e z_A é o escore-padrão do teste A.

Por exemplo:

Se a média da Forma A foi de 50 com um DP de 10 e a média da Forma B de 40 com DP de 8, então temos que a inclinação da curva da Forma A será 8/10 = 0,8 e o intercepto será 40 – (8/10)50 = 40 – 40 = 0,0. Qual será o escore em B correspondente a um escore de 30 em A?

Resposta: B = 0,8(30) + 0,0 = 24.

Isto significa que o escore de 30 na Forma A corresponde a um de 24 na Forma B, indicando que esta última é mais difícil que a Forma A. Com os valores da inclinação de A e do intercepto, você pode equiparar qualquer escore obtido na Forma A com os obtidos na Forma B; inclusive, você verá que a diferença em dificuldade entre as duas formas varia com diferentes escores. Cf. a tabela 9-2, onde é feita a equiparação entre vários escores da Forma A e da Forma B e onde se mostra que, com diferentes escores, se altera o nível de dificuldade entre as duas formas do teste.

Tabela 9-2. Equiparação de escores entre A e B e níveis de dificuldade

Forma A	Cálculo para B	Forma B	Dificuldade
90	0,8(90) + 0,0	72	90 – 72 = 18
80	0,8(80) + 0,0	64	80 – 64 = 16
70	0,8(70) + 0,0	56	70 – 56 = 14
50	0,8(50) + 0,0	40	50 – 40 = 10
30	0,8(30) + 0,0	24	30 – 24 = 6
10	0,8(10) + 0,0	8	10 – 8 = 2

Você vê na tabela que os escores da Forma B foram equiparados aos da Forma A e que o nível de dificuldade entre as duas formas varia com diferentes valores da escala (cf. coluna *Dificuldade*). Com os dados da tabela 9-2 você pode construir a linha de regressão dos escores brutos da Forma B sobre a Forma A. Cf. a figura 9.4, onde é mostrada também a linha de regressão da equiparação pela média (onde a Forma B tem 10 pontos de maior dificuldade que a Forma A).

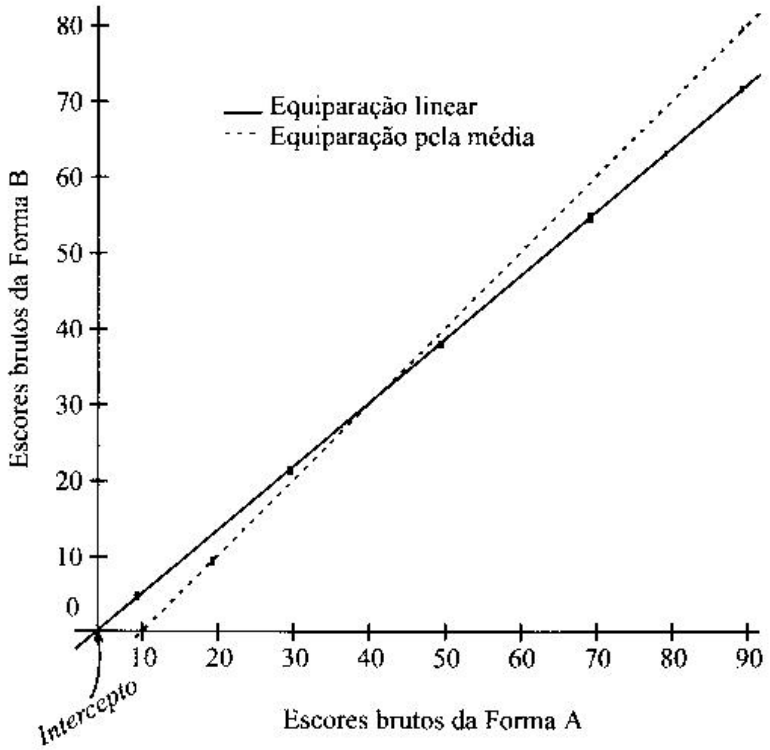


Figura 9-4. Linha de regressão de equiparação de escores de duas formas do teste

2.1.3 – O caso de teste de ancoragem

Em todos os casos considerados até agora, os grupos de comparação para fins de equiparação são randômicos, isto é, eles são equivalentes. No caso do delineamento com teste de ancoragem, os grupos de comparação já não são mais considerados equivalentes e, assim, os métodos de equiparação devem levar isto em conta. Relembre, primeiro, que no caso de teste de ancoragem temos a seguinte situação:

> dois testes ou formas do teste, A e B, que medem o mesmo traço latente, são aplicados a duas amostras diferentes e não equivalentes de sujeitos. Entretanto, há um subgrupo de itens X que entra tanto no teste A quanto no B. Assim, na verdade, temos três testes para levar em conta neste cômputo da equiparação dos escores de A e B, isto é, testes A, B e X.

Nesta situação, a equiparação dos escores pode ser feita através da regressão linear. Há uma série de métodos lineares (cf. Kolen & Brennan, 1995; Rabello, 2001); vamos aqui considerar o de Angoff (1984). Segundo este autor, o cálculo dos quatro parâmetros necessários para a análise de equivalência dos escores pelo método de transformação linear é o seguinte:

$$\overline{A} = \overline{A}_1 + b_{AX_1}\left(\overline{X} - \overline{X}_A\right)$$

$$\overline{B} = \overline{B}_2 + b_{BX_2}\left(\overline{X} - \overline{X}_B\right)$$

$$s_A = \sqrt{s_{A_1}^2 + b_{AX_1}^2\left(s_X^2 - s_{X_1}^2\right)}$$

$$s_B = \sqrt{s_{B_2}^2 + b_{BX_2}^2\left(s_X^2 - s_{X_2}^2\right)}$$

onde,

b_{AX_1} = regressão do teste A sobre os itens pertencentes a X na amostra 1 de sujeitos

b_{BX_2} = regressão do teste B sobre os itens pertencentes a X na amostra 2 de sujeitos

$\overline{X}$ = média global do teste X, isto é, incluindo os dados das duas amostras de sujeitos

$\overline{X}_1$ e $\overline{X}_2$ = a média do teste X respectivamente na amostra 1 e na amostra 2 de sujeitos

s^2_{A1} e s^2_{B2} = variância do teste A na amostra 1 e do teste B na amostra 2.

Obtidos estes valores, o cálculo segue a fórmula 9.4, ou seja,

$$B = \frac{s_B}{s_A} A + \left[M_B - \frac{s_B}{s_A} M_A \right].$$

O cálculo do erro padrão, neste caso, segue a seguinte fórmula:

$$s_E = \sqrt{\frac{2s_B^2 \left(1 - r^2\right) \left[\left(1 + r^2\right) z_A^2 + 2\right]}{N_t}}$$

onde,

$$r = \frac{b_{AX_1}}{s_A} + \frac{b_{BX_2}}{s_B}.$$

Nota: Para trabalhar esta fórmula de Angoff e inúmeras outras que existem para equiparação com teste de ancoragem, você deve consultar obras especializadas, como a de Kolen e Brennan (1995), Feldt e Brennan (1989) e Levine (1955).

2.2 – Equiparação equipercentílica (os percentis)

Esta é uma transformação não linear dos escores brutos e foi desenvolvida por Braun e Holland (1982); sua fórmula é complicada (9.7):

$$e_B = G^{-1} F_A \qquad (9.7)$$

onde,

e_B = a função de equiparação para converter escores da Forma A para a Forma B

G^{-1} = a inversa da distribuição cumulativa da função e_Y, sendo G a distribuição cumulativa da população da Forma B

F_A = a função da distribuição cumulativa dos escores de A na população.

Em vez de utilizar essa fórmula complexa, você pode utilizar simplesmente a transformação percentílica dos escores; ela consiste em se comparar, não os escores brutos, mas sim os percentis destes escores. A técnica é conhecida como o método equipercentílico. Por exemplo: o

sujeito recebeu um escore bruto de 50 no teste A e 60 no teste B. Obviamente os dois escores são diferentes, mas se soubermos que o escore 50 do teste A corresponde ao percentil 80 naquele teste e o escore 60 do teste B também corresponde ao percentil 80 neste teste, então ambos os escores são equivalentes, isto é, o sujeito recebeu percentil 80 em ambos os testes, o que vem a significar que receber um escore de 50 no teste A equivale a receber um de 60 no teste B. Veja o exemplo dos escores de 200 sujeitos respondendo a duas formas de 10 itens de um teste de aptidão; os dados estão na tabela 9-3. A figura 9-5 compara as percentagens acumuladas (os percentis) dos escores das duas formas.

Tabela 9-3. Equiparação percentílica dos escores de duas formas do teste

Escore bruto (T)	Forma A			Forma B		
	Frequência (f)	Percentagem (%)	Percent. acumulada	Frequência (f)	Percentagem (%)	Percent. acumulada
10	4	2	100	2	1	100
9	8	4	98	2	1	99
8	12	6	94	4	2	98
7	32	16	88	8	4	96
6	44	22	72	12	6	92
5	50	25	50	32	16	86
4	30	15	25	44	22	70
3	12	6	10	50	25	48
2	4	2	4	30	15	23
1	2	1	2	12	6	8
0	2	1	1	4	2	2

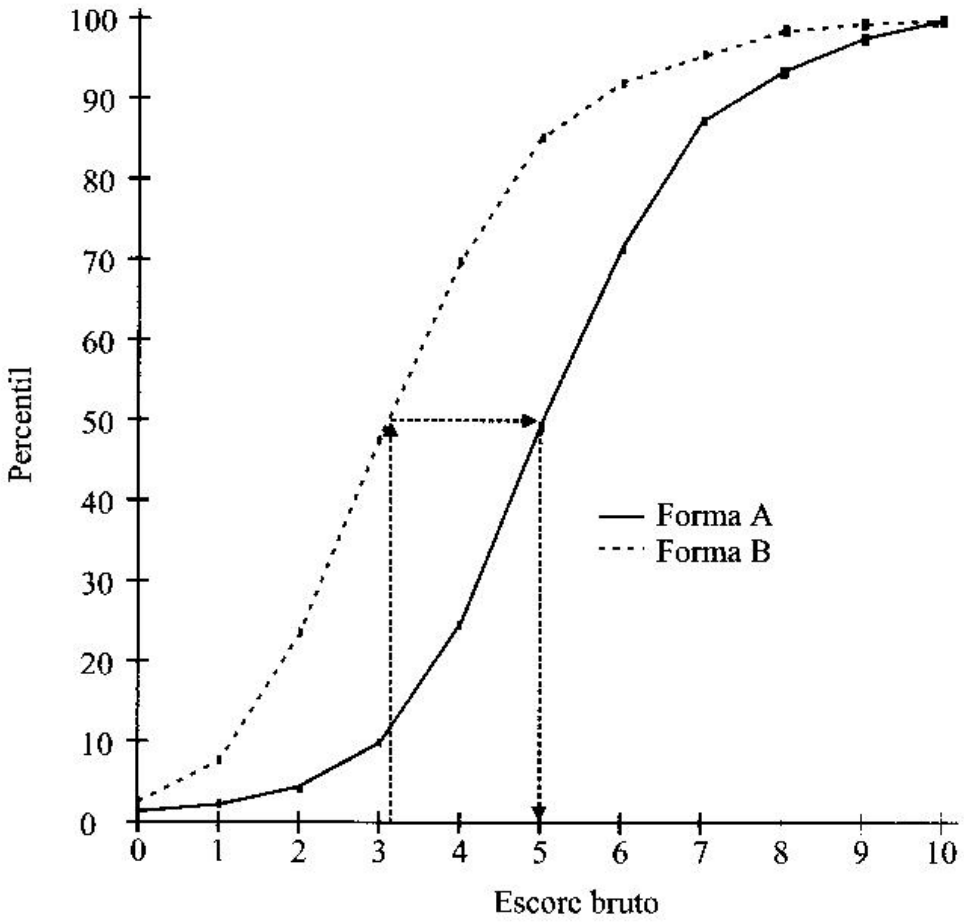


Figura 9-5. Correspondência percentílica dos escores de duas formas do teste

Assim, você vê na tabela 9-3 que um escore percentílico de 98 corresponde a um escore bruto de 9 na Forma A e de 8 na Forma B.

2.3 – Métodos da TRI

Também no caso da equiparação via TRI há uma série de métodos diferentes (Hambleton, Swaminathan & Rogers, 1991; Muñiz, 1990; Kolen & Brennan, 1995; Rabello, 2001), tais como regressão linear, método da média e do sigma, método robusto da média e do sigma, método da curva característica. Aqui iremos apenas ilustrar o método da regressão linear utilizado com o delineamento de teste de ancoragem.

O problema aqui é o seguinte: você aplicou dois testes paralelos ou duas formas equivalentes de um teste (A e B, por exemplo) a dois grupos não equivalentes de sujeitos, sendo que uma série de itens em ambos os testes são os mesmos (o teste de ancoragem). Se você fizer a análise dos parâmetros dos dois testes (de itens e dos tetas dos sujeitos), seguramente

você obterá valores diferentes para estes parâmetros nos dois testes, pelo menos devidos aos erros de medida. Esta ocorrência é bastante usual particularmente em duas situações, a saber: (1) quando você quer comparar as aptidões dos sujeitos baseadas em testes diferentes (que, obviamente, estejam medindo a mesma coisa) ou, (2) quando você quer incluir itens novos dentro de um banco de itens já calibrados. Nestes casos, você deve garantir que os parâmetros, tanto dos tetas (aptidão), quanto dos itens (dificuldade, discriminação, chute) se encontrem na mesma escala de medida. Para conseguir tal façanha, você deve fazer a equiparação (*equating*, *scaling*) de todos estes parâmetros para colocá-los todos numa mesma escala e, assim, poderem ser diretamente comparados.

Como fazer isso? Como dito acima, há vários métodos para resolver este problema. Contudo, observe as seguintes relações lineares que os parâmetros do teste A têm com os do teste B:

– Aptidão (teta): $\theta_B = d\theta_A + k$

– Dificuldade: $b_B = db_A + k$

– Discriminação: $a_B = \frac{a_A}{d}$

– Chute: $c_B = c_A$

onde, *d* e *k* são constantes que devem ser estimadas; os subscritos se referem aos testes A e B.

O cálculo de *d* e de *k* se faz da seguinte maneira:

$$d = \frac{s_{Bx}}{s_{Ax}}$$

$$k = M_{Ba} - dM_{Aa}$$

onde,

– s_{Bx} e s_{Ax} são os desvios-padrão no teste de ancoragem X para os sujeitos das amostras do teste A e B respectivamente;

– M_{Ba} e M_{Aa} são as médias no teste de ancoragem X para os sujeitos das amostras do teste A e B respectivamente.

Exemplo: Obviamente, os cálculos desses parâmetros e constantes devem ser deixados para programas de computador, tais como o BILOG. Contudo, um exemplo ilustrará como essas equações funcionam. Cf. a tabela 9-4.

Tabela 9-4. Estatísticas para equiparação com teste de ancoragem

Grupo	Escore	Média	DP (s)
1	A	15,82	6,50
1	X	5,10	2,35
2	B	18,60	6,80
2	X	5,80	2,45

Com os dados da tabela, descobrimos que:

$$d = \frac{s_{Bx}}{s_{Ax}} = 2,45/2,35 = 1,04$$

$k = M_{Ba} - dM_{Aa} = 5,80 - 5,10 = 0,70.$

Assim, a equiparação dos parâmetros dos dois testes A e B será a seguinte:

– Aptidão (teta): $\theta_B = d\theta_A + k = 1,04\theta_A + 0,70$

– Dificuldade: $b_B = db_A + k = 1,04b_A + 0,70$

– Discriminação: $a_B = \frac{a_A}{d} = a_A/1,04$

– Chute: $c_B = c_A = c_A$

Tendo os valores dos parâmetros em A, podemos equiparar os mesmos parâmetros para B. Cf. exemplos na tabela 9-5.

Tabela 9-5. Equiparação de parâmetros de dois testes via TRI

Parâmetros	Valores em A	Valores Equivalentes em B
a	0	0,00
	1	0,96
	2	1,92
	3	2,88
	4	3,85
b	-3	-2,42
	-2	-1,38
	-1	-0,34
	0	0,70
	1	1,74
	2	2,78
	3	3,82
θ	-3	-2,42
	-2	-1,38
	-1	-0,34
	0	0,70
	1	1,74
	2	2,78
	3	3,82

CAPÍTULO 10
Os testes psicológicos e o computador: Flexibilizando a aplicação dos testes

Modernamente, com a popularização do computador, vem se tornando cada vez mais prática e proveitosa a utilização deste aparelho na situação de testagem e no futuro ela será certamente a forma privilegiada no uso dos testes psicológicos, particularmente no caso da testagem dita adaptativa, do qual falaremos já em seguida. Nesta área já há uma plêiade de expressões que começam até a introduzir confusão. Anote algumas dessas expressões, as quais estão praticamente sempre atreladas ao uso do computador com testes psicológicos: testes sob medida, testes adaptados ao sujeito, testes de nível flexível, testes ramificados, testes individualizados, testes programados, testes sequenciais, além das expressões em inglês *tailored tests*, *adaptive testing*, *computerized adaptive testing*. Essa abundância de expressões indica que o campo parece um tanto minado e, por isso, é preciso esclarecer um pouco esta área e verificar o papel do computador no contexto dos testes psicológicos.

1 – Função do computador na testagem psicológica

Na verdade, é preciso distinguir, pelo menos, dois usos muito distintos que se fazem do computador na testagem psicológica: de fato, ele pode ser utilizado como *aplicador* de testes ou como *executor* de testes. Vejamos um pouco esta distinção.

1.1 – O computador como aplicador de testes (testes informatizados)

Quando se fala de testes informatizados no Brasil, praticamente se quer falar deste tipo de uso do computador com os testes. Este é o caso em que o computador substitui tanto o material utilizado (folheto do teste, folha de resposta, crivos de correção, etc.), quanto o próprio

psicólogo aplicador. Já é uma tarefa interessante do computador no contexto dos testes, porque ele pode:

- *Apresentar* muito melhor as questões do teste: maior qualidade do material, maior clareza, pode repetir sem se cansar as explicações, pode moldar desenhos, modificar cores, dar intervalos exatos de tempo, de distâncias, de intervalos, de níveis de estímulos, etc.;
- *Corrigir* as respostas sem errar e com muita rapidez;
- *Produzir o perfil* de respostas do sujeito e enquadrá-lo imediatamente nas tabelas de interpretação;
- Até *interpretar* o perfil psicológico do sujeito;
- *Produzir registros* legíveis, em número sem fim e transmiti-los à distância imediatamente;
- *Motivar* os testandos, porque estes normalmente se fascinam ao interagir com o computador, inclusive porque ele não se cansa de atender, não se irrita, não azucrina o testando, segue o ritmo normal do testando, deixa o testando se expressar livremente sem o constrangimento que um experimentador humano pode acarretar.

Então, você pode elencar algumas vantagens que o computador tem como aplicador de testes com respeito a um aplicador humano, o psicólogo, por exemplo. As seguintes são algumas vantagens ou possibilidades óbvias, que praticamente só com o computador se tornam viáveis:

1) Uso maior de *gráficos* como itens de teste em lugar de texto; este aspecto é particularmente importante, porque se sabe que a linguagem introduz nos testes fatores que se confundem com o que o teste de fato quer medir; com itens gráficos tal evento é menos provável. Ademais com itens gráficos você tem no computador uma gama de periféricos que podem ser utilizados para formatar os itens e as respostas a eles, tais como o cursor, o *joystick*, os ícones, e outros, como faz o sistema COMPASS (ACT, 1993);
2) Medida exata do *tempo de resposta*. O tempo de reação vem sendo estudado desde o começo da psicologia e vem sendo considerado um elemento relevante na avaliação das aptidões humanas, sobretudo sob o tópico de tempo de latência. O

computador obviamente é capaz de registrar tal comportamento com muito maior precisão do que faria um aplicador humano fazendo uso de um cronômetro (Thissen,1983);

3) O uso de *multimídia*. Esta área ainda não está muito presente na testagem psicológica, mas o computador certamente possibilita o uso de vídeos, sons e, até, da realidade virtual.

Obviamente, o computador ainda não substitui o aplicador humano na observação do comportamento do sujeito, quando isto for importante numa testagem. Também a interpretação do perfil psicológico é mais limitada do que a de um psicólogo humano. O computador é certamente muito superior ao aplicador humano naqueles aspectos do teste que são mais mecânicos, exigindo capacidade de memória de armazenamento, rapidez e precisão, e, quiçá, como redutor da ansiedade na tomada dos testes, mas é mais limitado na interpretação psicológica dos resultados de um teste.

1.2 – O computador como executor de testes (testagem adaptativa)

Como executor de testes, a função do computador é muito mais do que ser um simples aplicador de testes; aqui também ele faz isso, mas faz muito mais. De fato, o computador cria testes, adaptando-os a certos níveis de habilidade dos sujeitos (*optimal test assembly* – montagem de testes otimizados) ou cria o teste na hora para cada testando diferente, isto é, ele adapta a testagem a cada testando (*computerized adaptive testing* – testagem adaptativa). Como assim? Como essas duas tarefas do computador são, digamos, tarefas mais nobres, vamos tratá-las em separado a seguir. Entretanto, para que essas duas tarefas nobres possam ser realizadas pelo computador, há duas condições preliminares necessárias a serem cumpridas, a saber, banco de itens e algoritmo de sorteio dos itens. O algoritmo será discutido dentro das seções de montagem de testes otimizados (*optimal test assembly*) e dos testes sob medida (*computer adaptive testing*), mas temos primeiro que falar um pouco sobre o banco de itens.

2 – Banco de itens

Sobre a questão do banco de itens já falamos brevemente no capítulo 4, mas devemos aqui dizer mais algumas coisas. Continuando o pensamento do parágrafo anterior, devemos dizer que o computador, na verdade, não cria coisa nenhuma, nem testes nem itens. Ele pode analisar

a qualidade psicométrica de um número sem fim de itens e incluí-los num armazém, onde estes têm sua carta individual de identidade. Tendo este arsenal de itens, aí sim o computador pode criar testes sem fim; quanto maior for o tamanho deste arsenal, chamado de *banco de itens*, mais testes o computador pode criar. Agora, quem cria os itens é o psicólogo pesquisador e é ele que vai coletar a informação empírica sobre cada item através da pesquisa científica. O computador analisa esta informação e produz a carteira de identidade do item. Esta carteira contém os dados (ditos parâmetros), discutidos no capítulo 5, tais como: que traço psicológico o item mede, qual seu índice de dificuldade, qual seu índice de discriminação, qual a percentagem de chute que ele provoca, se o item tem sentidos diferentes para diferentes sujeitos ou populações, qual o nível de informação que ele traz (curva de informação do item) e alguns dados mais. No futuro, a grande riqueza daqueles que trabalham com testes são precisamente estes bancos de itens. A sua construção exige pesquisa científica laboriosa e demorada (fala-se de que é necessário trabalhar, pelo menos, três anos para construir um único banco de itens), pois se trata de criar centenas e milhares de itens para cada processo psicológico e demonstrar a sua qualidade psicométrica (calibração). Além disso, é preciso fazer a manutenção periódica dos bancos de itens, porque, com o tempo, eles podem perder suas características de bons itens, isto é, sua validade. Como é que se faz isto? Pela técnica moderna da Psicometria, chamada Teoria de Resposta ao Item (TRI), esta tarefa é facilitada, porque ela permite estabelecer os parâmetros dos itens que falamos acima independentemente da amostra de sujeitos utilizada; daí é possível incluir sempre novos itens diretamente comparáveis com os já inclusos no banco, com cada pesquisa que se faz sobre o processo psicológico em questão. A técnica para esta façanha, entre outras, consiste em aplicar os novos itens juntamente com uma amostra de itens já incluídos no banco a uma amostra razoável de sujeitos e estimar os parâmetros dos novos itens em confronto com os dos itens utilizados do banco de itens (cf. capítulo 9 sobre equiparação). Assim os novos itens entram no banco nas mesmas condições que os velhos, sendo que os velhos ficam assim também revalidados (sobre banco de itens cf. van der Linden, 1986; Applied Psychological Measurement, 1986; de Gruijter & van der Kamp, 1976; Chopin, 1976; Millman & Arter, 1984; Wright & Bell, 1984).

3 – Montagem de testes otimizados (*optimal test assembly*)

Se você dispõe de um banco volumoso de itens (digamos, 500 a 1000 itens) calibrados (que possuem os parâmetros discutidos no capítulo 5), então você pode utilizar o computador para construir muitos testes diferentes para avaliar o traço latente que o banco de itens cobre. Os testes clássicos eram fixos e definidos uma vez por todas e tinham a preocupação de cobrir equitativamente toda a extensão do traço que queriam medir. Tipicamente eles tinham itens desde os mais fáceis aos mais difíceis, com predomínio de itens de dificuldade mediana, já que queriam ser úteis para a população em geral. Isto é, eles queriam cobrir todos os níveis de teta de um dado traço latente. Com essa preocupação, esses testes se mostravam adequados para avaliar sujeitos que tinham níveis medianos de θ e avaliavam menos bem sujeitos com θ baixos ou altos. Isso era o caso, porque você se lembra do capítulo 4 que a TCT se interessava em construir testes e, uma vez construídos, eles eram intocáveis e herméticos. Assim quando você queria utilizar um dado teste desse tipo para fins específicos como, por exemplo, selecionar os 30% melhores num dado traço latente, você devia lançar mão do teste que foi construído para avaliar toda a extensão do θ deste traço, teste que era bom para θ medianos e ruim para θ extremos. Como, nesse exemplo, você queria precisamente medir θ extremos, então o teste disponível se apresentava precário.

Esta situação modificou-se com a TRI, porque esta está interessada em construir e calibrar itens e não testes. Assim, tendo um banco de itens calibrados disponíveis, você pode construir quantos testes você quiser ou o banco de itens permitir, orientando estes testes para os fins específicos que você tenha em mente, como o caso mencionado na seleção dos 30% melhores sujeitos no traço latente. Assim, através do computador, você pode construir testes para qualquer nível de θ que queira; basta definir o objetivo e as restrições que você quer impor ao teste a ser construído pelo computador. Essa técnica de construir testes chama-se *montagem de testes*, porque de fato o computador não está construindo testes, ele apenas monta os itens para compor um teste, seguindo as instruções e as restrições impostas por você, como numa linha de montagem. Já em 1968, Birnbaum percebeu a relevância e utilidade de tal técnica, mas a falta de computadores tornava sua ideia difícil de ser viabilizada na prática. Cf. a figura 10-1 para visualizar o que acabamos de falar, isto é, montar testes para diferentes níveis de θ, isto é, para diferentes grupos de sujeitos.

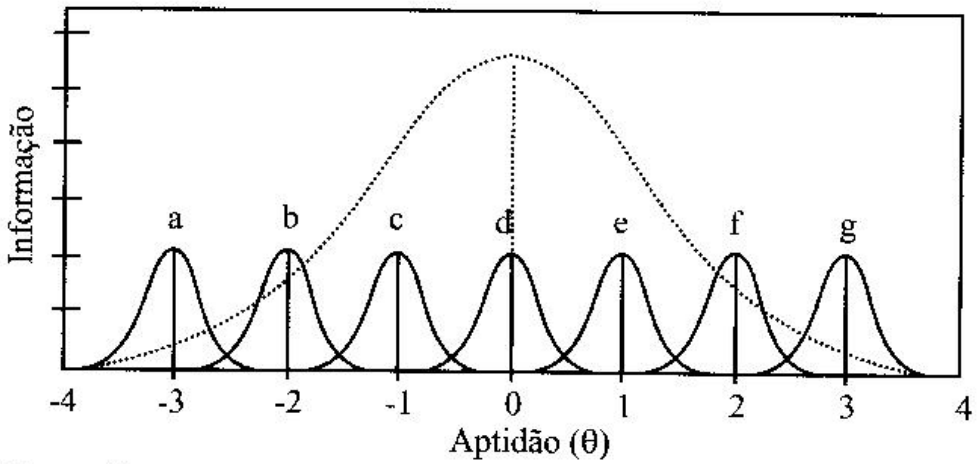


Figura 10-1. Testes montados para diferentes valores de θ

A figura 10-1 mostra a montagem de sete testes diferentes (*a* até *g*) medindo o mesmo traço latente, mas para níveis diversos de θ: para cada nível foi montado um teste diferente, de sorte que o teste que serve para avaliar um θ = 2 não serve para avaliar um θ = -1. A curva pontilhada indica como a TCT construiria um teste único para avaliar todos os níveis do θ, mas que de fato avalia adequadamente apenas os níveis medianos desse traço. Você também vê na figura que os testes montados apresentam maior nível de informação sobre o θ no ponto central do respectivo teste, caindo essa informação para as caudas das curvas; de sorte que, por exemplo, o teste montado para avaliar θ = 0 avalia melhor a faixa de θ que se situa entre -0,4 e +0,4.

Para mandar o computador montar testes desta natureza você tem que dar a ele instruções, indicando os objetivos e as restrições que você quer que sejam atendidos. Por exemplo, uma série de tais instruções e restrições poderia ser a seguinte: monte um teste para avaliar o θ tal que:

1) a dificuldade dos itens se situe entre -1 e +1;

2) a discriminação dos itens não seja inferior a 0,95;

3) o chute não ultrapasse os 10%;

4) o teste tenha 15 itens;

5) os itens não sejam repetidos;

6) 40% dos itens se situem entre os níveis de dificuldade (b) -0,4 e +0,4.

Com tais instruções, o computador é capaz de montar uma série de testes equivalentes para avaliar o que você quer; obviamente, se o banco de itens for suficientemente grande.

Técnicas mais exatas para montar tais testes podem ser encontradas em van der Linden (1992, 1993, 1995), Adema (1990, 1992), Adema e van der Linden (1989), Armstrong e Jones (1992), Armstrong, Jones e Wu (1992), Baker, Cohen e Barmish (1988), Boekkooi-Timminga (1987, 1989, 1990), de Gruijter (1990), Luecht e Hirsch (1992), van der Linden e Boekkooi-Timminga (1988, 1989).

4 – Testes sob medida (*computerized adaptive testing – CAT*)

Levando essa ideia da montagem de testes mais adiante, pode-se perguntar: por que montar testes para faixas de θ, se o computador pode montar um teste específico para cada sujeito? É o que faz a técnica dita de testes sob medida ou testes adaptados ao sujeito (*tailored testing* ou *computer adaptive testing* – CAT).

Esta técnica recebe o nome de testes adaptados ao sujeito precisamente porque ela visa aplicar no sujeito tarefas de dificuldade que se aproxima do nível de aptidão do sujeito, de cada sujeito. Para conseguir esta façanha, é preciso inicialmente que se conheça o nível de aptidão do sujeito para, em seguida, aplicar os itens de dificuldade correspondente a este nível de habilidade do sujeito. Mas aqui surge algo de curioso: para montar um teste adequado para o sujeito, devo primeiro conhecer o nível de aptidão do sujeito; mas se conheço o nível de aptidão do sujeito para que serve montar um teste para avaliar a aptidão do mesmo sujeito? Este dilema, chamado de paradoxo de delineamentos de testes, foi resolvido de várias formas pela teoria dos testes. Vamos ver como a TRI o resolveu.

Desde a Psicometria Clássica (TCT) se sabia que um teste traz maior informação sobre o sujeito se ele se adapta ao nível de habilidade deste; isto é, de alguma maneira é preciso saber mais ou menos e preliminarmente o nível de aptidão do sujeito para se aplicar um teste adequado a ele. Desde os tempos de Lord (1980), dois enfoques são utilizados para resolver este problema dentro da TRI, a saber, o enfoque de nível duplo (estágio duplo; Cronbach & Gleser, 1965) e enfoque de níveis múltiplos (estágios múltiplos, multinível) ou níveis flexíveis. Vamos ver o que esses enfoques representam.

No enfoque de *estágio duplo* há dois passos no processo de avaliação, a saber:

1) aplica-se a todos os sujeitos um mesmo teste, um teste de triagem. O escore neste teste vai dar uma estimativa do nível de aptidão dos sujeitos;

2) em função desse escore, é aplicado um segundo teste (teste principal) ao sujeito, agora um teste adaptado ao nível de aptidão do mesmo.

O problema ou a dúvida que esta técnica levanta consiste em decidir qual deve ser o tamanho do teste de triagem, para poder dar uma informação confiável e, ao mesmo tempo, não encher a paciência dos sujeitos, uma vez que vão se submeter a outro teste.

No enfoque de *estágios múltiplos* vai se aplicando ao sujeito um item depois do outro, até que se tenha uma estimativa satisfatória do nível de aptidão do sujeito, sem ter que aplicar um teste de triagem. Os problemas com esta técnica são: com que item iniciar, como escolher o item seguinte e quando parar?

Esta técnica dos níveis múltiplos é a preferida dos psicometristas, porque com pouco tempo e poucos itens produz a mesma informação sobre os sujeitos que o uso de testes longos e cansativos. De fato, para obter o mesmo resultado, com a técnica dos testes adaptativos você precisa cerca de 40% menos itens e 50% menos tempo de testagem. Contudo, para implementar praticamente esta técnica, você tem que resolver alguns problemas, obedecendo as quatro regras seguintes (van der Linden, 1995; Hambleton, Zaal, & Pieters, 1991; Reckase, 1983; Kingsbury & Weiss, 1983): você precisa decidir uma regra para

1) estimar a habilidade do sujeito a partir das respostas por ele dadas aos itens que precederam;

2) escolher o primeiro item da série;

3) escolher o próximo item a ser apresentado ao sujeito em função da sua habilidade até ali estimada ou reestimada;

4) terminar a testagem.

4.1 – Regras para estimar a habilidade

Os métodos para estimar a habilidade são a máxima verossimilhança (Weiss, 1982) e os de Bayes (Owen, 1975; Hambleton, Swaminathan & Rogers, 1991). Esses métodos funcionam mais ou menos da seguinte maneira:

1) seleciona-se um item de dificuldade média. Se o sujeito acertou o item, ainda não dá para utilizar a máxima verossimilhança (MV) até que o sujeito tenha respondido errado pelo menos um item. Assim, aplica-se um item mais difícil. Se o sujeito acertar, aplica-se um mais difícil ainda. Se ele errar, aí já dá para utilizar a MV;

2) na verdade, a esta altura as respostas do sujeito produziram um padrão, que é o seguinte: acertou, acertou, errou; ou seja, 1 1 0. Com este padrão, a MTV pode calcular uma primeira aproximação de estimativa da habilidade do sujeito. Suponha que essa estimativa deu um teta de 0,5;

3) assim, sabemos que os itens restantes a serem aplicados serão itens cuja dificuldade *b* gira em torno de 0,5;

4) a cada novo item introduzido, a máxima verossimilhança irá reestimar o teta do sujeito, até que a regra de término da sessão seja acionada.

4.2 – Regras para escolher o primeiro item

A regra que dita a escolha do primeiro item é a seguinte: escolha aquele item que produz a maior informação para a habilidade da população à qual o sujeito pertence. Normalmente isto significa escolher um item de dificuldade mediana. O que fica de incógnito nesta regra é saber a habilidade média da população donde o sujeito saiu. Para resolver este problema, pode-se recorrer a informações de fontes externas ao teste, tais como escores em testes anteriores, opinião de professores para uma população acadêmica, desempenho na profissão, e outros. A escolha do primeiro item é importante, porque quanto mais próximo ele for do nível de habilidade do sujeito, mais rápido se chega à estimação final do seu nível de habilidade. Quanto mais longínquo da habilidade do sujeito for o item escolhido, mais iterações são necessárias para se chegar à estimação do verdadeiro nível de aptidão do sujeito, perdendo-se, assim, rapida-

mente as vantagens de uma testagem adaptativa, além de acumular as chances de aumento de erros na estimação do θ.

4.3 – Regras para escolher o próximo item

Conhecida a habilidade inicial do sujeito (no caso o θ = 0,5), os itens que vão ser utilizados com este sujeito terão a dificuldade em torno de 0,5. Assim, o próximo item a ser apresentado ao sujeito deve ser aquele que, dentro da faixa de habilidade do sujeito estimada no passo anterior, produz a maior informação sobre o nível de θ do sujeito (informe dado pela curva de informação dos itens que se encontram nesta faixa de habilidade do sujeito). Isto significa o seguinte: para um teta de 0,5, o item com maior nível de informação foi, digamos, o item 5 do banco de itens; então, este item é aplicado ao sujeito. Ele acerta e seu θ será reestimado e resulta em θ = 0,6. Para θ = 0,6, o item que produz a maior informação é, digamos, o item 120; então, este item é aplicado ao sujeito, e etc., até que entre em vigor a regra de terminar a sessão.

4.4 – Regras para terminar a testagem

A testagem adaptativa pode continuar até esgotar o banco de itens pertinentes à faixa de habilidade do sujeito, se não se estabelece uma regra de término. Esta regra é definida em função do padrão de resposta do sujeito. Este padrão logo começa a aparecer como uma alternância entre acertos e erros, já que se o sujeito acerta o item, um item mais difícil é apresentado; se ele erra este item, um item mais fácil é apresentado, o qual ele acerta, e assim por diante. De sorte que você vai ter um padrão do tipo 1 1 1 0 0 1 0 1 0 1, o qual pode ser examinado pela máxima verossimilhança. Manda-se parar o procedimento quando um certo nível predefinido de erro padrão é atingido.

Infelizmente não há muitos softwares disponíveis no mercado para trabalhar com testes sob medida. Em nível internacional, já estão aparecendo vários softwares que permitem a utilização da técnica dos testes adaptativos. Além do conhecido MicroCat da *Assessment Systems Corporation* (cf. uma revisão dele em Stone, 1989), vários outros estão apresentados no apêndice C.

CAPÍTULO 11
Introdução à análise fatorial[1]

1 – Introdução

A análise fatorial compreende uma série de técnicas estatísticas que trabalham com análises multivariadas e matrizes. Ela constitui uma técnica estatística imprescindível no contexto da Psicometria, sobretudo para a problemática da validação de instrumentos psicológicos. Ela tem muita coisa a dizer a respeito tanto da validade quanto da fidedignidade destes instrumentos. No caso da fidedignidade, por exemplo, é quase somente através dela que se pode estabelecer este parâmetro psicométrico para baterias de testes. Razão pela qual se faz necessária uma exposição sobre esta técnica. Embora ela seja matematicamente de uma complexidade demasiada, a presente exposição procurará contornar o mais possível os meandros matemáticos e explicitar a lógica e o modelo da análise fatorial para leitores não sofisticados em matemática.

2 – O modelo da análise fatorial

A análise fatorial é uma técnica estatística calcada sobre o pressuposto de que uma série de variáveis observadas, medidas, chamadas de variáveis empíricas ou observáveis pode ser explicada por um número menor de variáveis hipotéticas, não observáveis, chamadas precisamente de variáveis hipotéticas ou variáveis-fonte, mais conhecidas sob o nome de fatores. Estas variáveis-fonte seriam a causa do fato de que as variáveis observáveis se relacionam entre si, isto é, são responsáveis pelas intercorrelações (covariância) entre estas variáveis. Supõe-se que, se as variáveis empíricas se relacionam entre si, é porque elas têm uma causa comum que produz esta correlação entre elas. É a esta causa comum que se chama de

1. Uma exposição mais técnica pode ser encontrada no livro do mesmo autor *Análise Fatorial para Pesquisadores*, Petrópolis, RJ: Editora Vozes, s.d.

fator e cuja descoberta é precisamente a tarefa da análise fatorial. Então, nestas afirmações já fizemos dois postulados que a análise fatorial assume; (1) um número menor de variáveis-fonte é suficiente para explicar uma série maior de variáveis observáveis (princípio de parcimônia ou de *rank reduction*, isto é, redução do posto da matriz das intercorrelações entre as variáveis observáveis); (2) as variáveis-fonte são a causa da covariância entre as variáveis observáveis (princípio da causalidade – a análise fatorial é um modelo causal).

Podemos, assim, expressar o modelo da análise fatorial, dizendo que as variáveis observáveis são função de variáveis hipotéticas e ilustrar como na figura 11-1.

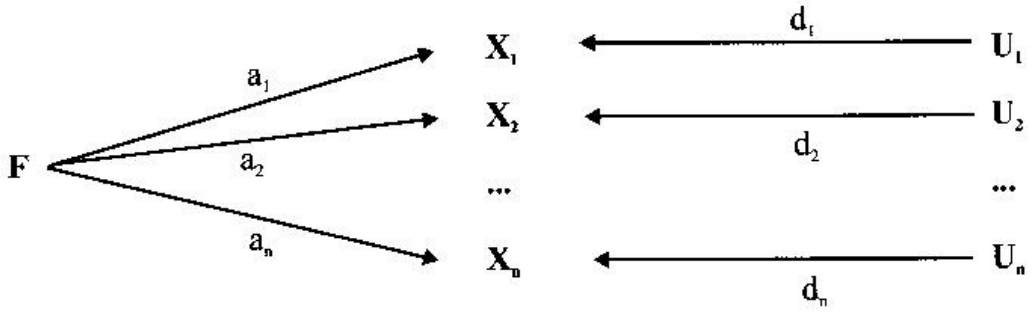


Figura 11-1. Representação do modelo fatorial

onde,

X_1 e X_2 são as variáveis observáveis 1 e 2

F é a variável-fonte comum às duas variáveis observáveis; é chamada de fator comum, já que é comum a mais de uma variável observável

U_1 e U_2 são as variáveis-fonte específicas (exclusivas) de cada variável observável; são chamadas de fatores únicos, já que são únicos ou específicos para cada variável observável

a_1, a_2, d_1, d_2 são os pesos (cargas) das variáveis observáveis nas variáveis-fonte.

Assim, este modelo nos permite afirmar que

– X_1 é igual à soma ponderada (por a_1 e d_1) de F e U_1

– X_2 é igual à soma ponderada (por a_2 e d_2) de F e U_2

Destas afirmações surgem as seguintes equações para as variáveis observáveis:

$$X_1 = a_1F + d_1U_1$$
$$X_2 = a_2F + d_2U_2 \qquad (11.1)$$

Observe que aqui fizemos mais um postulado, a saber, que a análise fatorial trabalha com equações lineares. Além disso, afirma-se que

$$\begin{aligned} &Cov(F, U_1) = 0 \\ &Cov(F, U_2) = 0 \\ &Cov(U_1, U_2) = 0 \end{aligned} \qquad (11.2)$$

A covariância entre os fatores comuns (F) e únicos (U_1, U_2) e estes entre si não existe. Isto é consequência do próprio modelo que distingue entre variáveis-fonte comuns, que podem ou não estar relacionadas entre si, e fatores únicos que, por definição, não podem estar relacionados aos fatores comuns e nem entre si; do contrário, não seriam fatores específicos de cada variável observável.

Assim, a equação geral que expressa a relação entre todas as variáveis (fonte comum, fonte específica e observável) será a seguinte:

$$X_j = a_jF_1 + a_jF_2 + \ldots\ldots\ldots + a_jF_r + d_jU_j \qquad (11.3)$$

onde,

j = 1, 2, ..., n (são as *n* variáveis observáveis de um teste, por exemplo)

r = 1, 2, ..., p (é o número de fatores comuns presentes no teste).

Exemplo: Cf. figura 11-2.

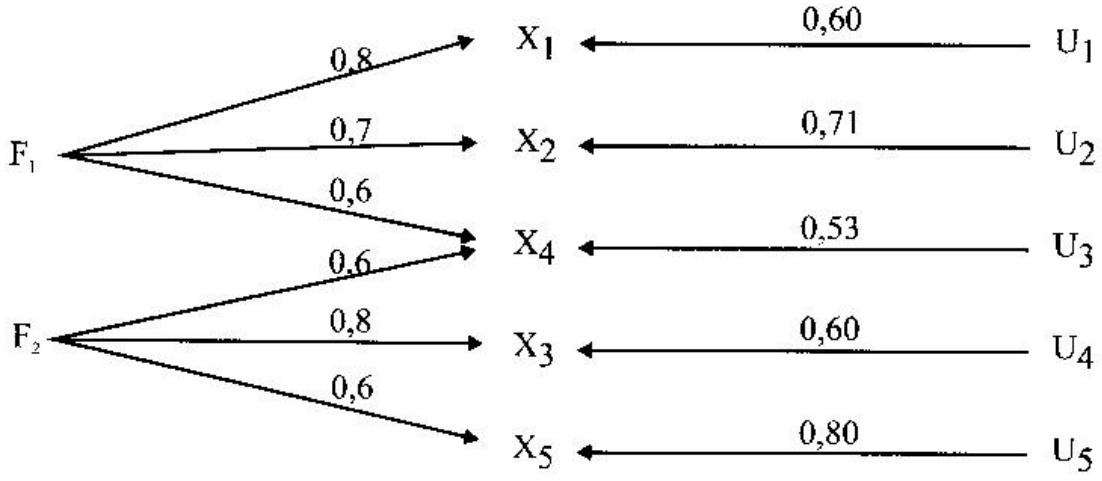


Figura 11-2. Modelo fatorial para 5 variáveis em 2 fatores comuns

Este exemplo mostra que a explicação das variáveis observáveis em função das suas variáveis-fonte é a seguinte:

$$X_1 = 0{,}8F_1 + 0{,}60U_1$$
$$X_2 = 0{,}7F_1 + 0{,}71U_2$$
$$X_3 = 0{,}6F_1 + 0{,}6F_2 + 0{,}53U_3$$
$$X_4 = 0{,}8F_2 + 0{,}60U_4$$
$$X_5 = 0{,}6F_2 + 0{,}80U_5$$

3 – Propriedades das variáveis observáveis em termos dos fatores

Se o modelo da análise fatorial for adequado, deve ser possível se poder estimar as propriedades estatísticas básicas das variáveis observáveis a partir do conhecimento dos parâmetros de suas variáveis-fonte. Na verdade, as variáveis observáveis podem ser, do ponto de vista da Estatística, exaustivamente descritas através de alguns parâmetros fundamentais: (1) individualmente, as variáveis observáveis podem ser descritas em termos da média e da variância; (2) em grupo, elas podem ser descritas em termos da covariância (a correlação). Assim, o modelo da análise fatorial deve ser capaz de produzir estes mesmos parâmetros para as variáveis observáveis.

Em termos de dados empíricos, os parâmetros estatísticos das variáveis observáveis são expressos como segue:

Média:

$$\overline{X} = \frac{\sum X_i}{N} \qquad (i = 1, 2, 3, \ldots\ldots, N \text{ sujeitos}) \qquad (11.4)$$
$$= E(X)$$

A média é a expectativa (abreviada como E) de X, isto é, o resultado esperado de X se este fosse medido um número infinito de vezes; de fato, neste caso, os erros cometidos de superavaliação e de subavaliação na medida do valor real de X se anulariam mutuamente, resultando que finalmente a medida seria do valor verdadeiro de X.

Variância:

$$\text{Var}(X) = \frac{\sum (X - \overline{X})^2}{N} = s_X^2, \text{ ou seja,} \qquad (11.5)$$

$$\text{Var}(X) = \frac{\sum [X - E(X)]^2}{N}$$
$$= E[X - E(X)]^2$$

Sabe-se que a variância são os desvios quadráticos médios, ou seja, a expectativa da média do somatório do quadrado dos desvios dos escores empíricos em relação ao escore médio desses escores.

Além disso, os escores empíricos podem ser expressos, sem perda de informação, em termos de escores padrões, isto é, da curva normal. Neste caso, os escores têm Média = 0 e Variância = 1. Assim, a equação da variância se reduz a $s_X^2 = E(X)^2$.

Covariância:

$$\text{Cov}(X, Y) = \frac{\sum [(X - \overline{X})(Y - \overline{Y})]}{N} = \frac{\sum xy}{N} \qquad (11.6)$$
$$= E[(X - \overline{X})(Y - \overline{Y})]$$
$$= E(XY)$$

A fórmula se reduz à última expressão, porque as médias de X e de Y, em termos de escore padrão, são 0.

Correlação:

$$r_{XY} = \text{Cov}(X, Y) = E(XY) \qquad (11.7)$$

A correlação, então, é a covariância entre as duas variáveis observáveis, X e Y. Na instância em que estas duas variáveis forem independentes, isto é, que não estejam correlacionadas, a covariância entre elas é nula, isto é, Cov(X,Y) = 0. Caso contrário, a covariância (a correlação) entre as variáveis varia de -1 a +1, seguindo a equação linear em que Y é expresso em termos de X, ou seja,

$Y = a + bX$.

Em conclusão, nós temos os seguintes parâmetros das variáveis:

Média: $\overline{X} = E(X)$

Variância: $s_X^2 = E(X)^2$

Covariância: $\text{Cov}(X,Y) = E(XY)$

Correlação: $r_{XY} = \text{Cov}(X,Y) = E(XY)$.

Para expressarmos estes parâmetros das variáveis observáveis em termos de fatores, basta substituir estas variáveis pelas suas equivalentes variáveis-fonte ou fatores expressos na equação 11.3. Vejamos os casos da variância e da covariância; tenha em sua frente a figura 11-1 para visualizar o que segue.

Variância de X em termos de fatores:

A variância da variável observável X_1 vem expressa por

$$\text{Var}(X_1) = s^2_{X_1} = E(X_1)^2, \qquad \text{pela equação 11.5}$$

X_1, por sua vez, vem expresso em termos de fatores como segue:

$$X_1 = a_1F + d_1U_1$$

Substituindo esta equivalência fatorial de X_1 na equação da variância, temos

$$s^2_{X_1} = E(a_1F + d_1U_1)^2 \qquad (11.8)$$

Desdobrando, esta equação resulta em

$$s^2_{X_1} = E\left(a_1^2F^2 + d_1^2U_1^2 + 2a_1d_1FU_1\right)$$

Evidenciando as constantes, a equação fica

$$s^2_{X_1} = a_1^2E(F^2) + d_1^2E(U_1^2) + 2a_1d_1E(FU_1)$$

Mas, as expectativas de F ao quadrado e de U ao quadrado são as respectivas variâncias, segundo a equação 11.5; igualmente, a expectativa de F e U juntos é a covariância entre estas duas variáveis, segundo a equação 11.6. Assim, a equação acima se expressa como:

$$s^2_{X_1} = a_1^2\text{Var}(F) + d_1^2\text{Var}(U_1) + 2a_1d_1\text{Cov}(X, U_1).$$

Entretanto, em termos padrões, sabemos que a variância é igual a 1 e, segundo a equação 11.2, sabemos que a covariância entre fatores comum e únicos é nula. Desta forma, a equação da variância de X_1 se reduz a

$$s^2_{X_1} = a_1^2 + d_1^2 = 1. \qquad (11.9)$$

Então, a variância da variável observável é a soma dos quadrados das suas cargas no(s) fator(es) comum(ns) e no fator específico dela.

Similarmente ocorre com X_2 onde a equação será

$$s^2_{X_2} = a^2_2 + d^2_2 = 1.$$

Tornando geral, esta fórmula fica:

$$s^2_j = a^2_{j1} + a^2_{j2} + \ldots\ldots\ldots + a^2_{jr} + d^2_j \qquad \begin{matrix} (j = 1, 2, \ldots\ldots, \text{n variáveis}) \\ (r = 1, 2, \ldots\ldots, \text{r fatores}) \end{matrix} \qquad (11.10)$$

onde, a_{j1} até a_{jr} representam as cargas dos *j* itens nos *r* fatores comuns às variáveis observáveis e d_j, as cargas destas variáveis nos seus respectivos fatores específicos.

Raciocínio similar pode ser empregado no caso da *covariância* das variáveis observáveis expressa em termos dos fatores. Vejamos:

Covariância entre X_1 e F:

Da equação 11.6, sabemos que a covariância entre X e Y é

Cov(X,Y) = E(XY).

Similarmente, a covariância entre X_1 e F será

$Cov(X_1,F) = E(X_1F)$.

Substituindo X_1 pelos seus equivalentes termos fatoriais, temos

$Cov(X_1,F) = E[(a_1F + d_1U_1)F]$, que efetuando dá (11.11)

$Cov(X_1,F) = E(a_1F^2 + d_1FU_1)$; evidenciando as constantes, temos

$Cov(X_1,F) = a_1E(F^2) + d_1E(FU_1)$, ou, em outras palavras,

$Cov(X_1,F) = a_1Var(F) + d_1Cov(FU_1)$

Mas como Var(F) = 1 e $Cov(FU_1) = 0$, em termos padrões, temos que

$$Cov(X_1,F) = a_1. \qquad (11.12)$$

Assim, a covariância entre a variável observável e os fatores é a carga desta variável no(s) fator(es) comum(ns). Esta carga é a covariância, a correlação bivariada e também a correlação parcial (beta) entre a variável e o fator comum, isto é,

$$a_1 = r_{X_1F} = \beta_1. \tag{11.13}$$

Similarmente, pode-se mostrar que

$$\text{Cov}(X_2, F) = a_2 = r_{X_2F} = \beta_2$$
$$\text{Cov}(X_1, U_1) = d_1 = r_{X_1U_1}$$

A *covariância entre as próprias variáveis observáveis* em termos dos fatores virá, da mesma forma, expressa como segue:

Pela equação 11.6, sabemos que Cov(X,Y) = E(XY). Então a covariância entre X_1 e X_2 será $\text{Cov}(X_1,X_2)$.

Expressando tanto o X_1 quanto o X_2 em termos de seus correspondentes fatores, temos

$$X_1 = a_1F + d_1U_1$$
$$X_2 = a_2F + d_2U_2$$

Substituindo estas equivalências dos X na equação da covariância, temos

$$\text{Cov}(X_1,X_2) = E[(a_1F + d_1U_1)\ (a_2F + d_2U_2)] \tag{11.14}$$

Expandindo, chegamos a

$$\text{Cov}(X_1, X_2) = E(a_1a_2F^2 + a_1d_2FU_2 + a_2d_1FU_1 + d_1d_2U_1U_2)$$

Evidenciando, temos

$$\text{Cov}(X_1, X_2) = a_1a_2E(F^2) + a_1d_2E(FU_2) + a_2d_1E(FU_1) + d_1d_2E(U_1U_2).$$

Isto é,

$$\text{Cov}(X_1, X_2) = a_1a_2\text{Var}(F) + a_1d_2\text{Cov}(FU_2) + a_2d_1\text{Cov}(FU_1) + d_1d_2\text{Cov}(U_1U_2)$$

Mas sendo a variância igual a 1 e as covariâncias entre fatores comum e específicos nulas, temos

$$\text{Cov}(X_1, X_2) = a_1 a_2 . \qquad (11.15)$$

Assim, a covariância entre duas variáveis observáveis em termos de fatores é expressa pelo produto de suas cargas fatoriais, que, aliás, também é a correlação entre elas, a saber,

$$a_1 a_2 = r_{X_1 X_2} = \beta_1 \beta_2 .$$

Assim, em conclusão, temos os parâmetros estatísticos das variáveis observáveis expressos em termos de fatores como segue:

Variância: $s^2_{X_1} = a_1^2 + d_1^2 = 1$

Covariância (= Correlação):

$$\text{Cov}(X_1, F) = a_1 = r_{X_1 F} = \beta_1$$

$$\text{Cov}(X_2, F) = a_2 = r_{X_2 F} = \beta_2$$

$$\text{Cov}(X_1, U_1) = d_1 = r_{X_1 U_1}$$

$$\text{Cov}(X_1, X_2) = a_1 a_2 = r_{X_1 X_2} = \beta_1 \beta_2 .$$

4 – Componentes fatoriais da variância

A variância das variáveis observáveis em termos de fatores pode ser dividida em várias porções, como ilustrado na figura 11-3.

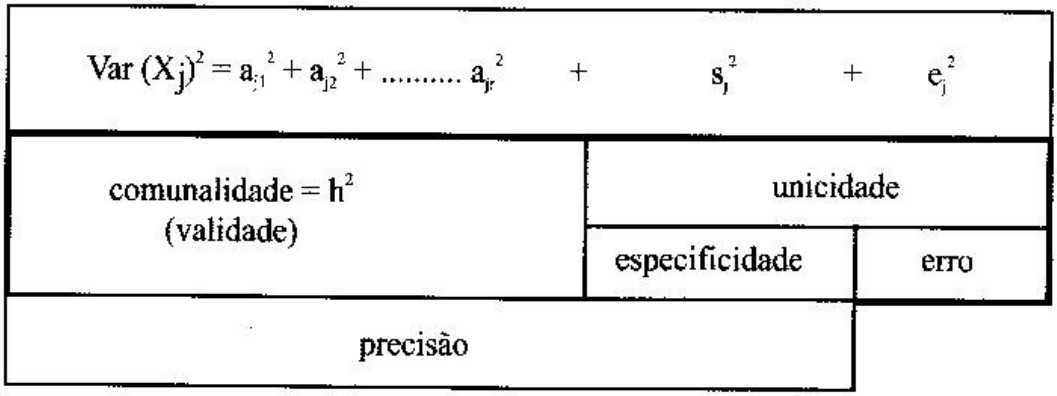


Figura 11-3. Distribuição dos componentes da variância

A variância pode ser decomposta de várias maneiras:

1) variância comum (comunalidade) e unicidade (especificidade + erro)

2) variância comum, variância específica, variância erro.

3) variância verdadeira (comunalidade + especificidade) e erro

Há uma série longa de equivalências que podemos fazer entre estes três componentes da variância. Por exemplo, sabemos que a variância total de uma variável é igual a 1. Assim vale para cada variável o que segue:

$$1 = h_j^2 + s_j^2 + e_j^2$$
$$h_j^2 = 1 - \left(s_j^2 + e_j^2\right)$$
$$r_{jj} = h_j^2 + s_j^2 \text{ (precisão)}$$
$$= 1 - e_j^2$$
$$h_j^2 = r_{jj} - s_j^2$$
$$s_j^2 = r_{jj} - h_j^2$$
$$u_j^2 = s_j^2 + e_j^2$$
$$= 1 - h_j^2$$
$$e_j^2 = 1 - \left(h_j^2 + s_j^2\right)$$
$$= 1 - r_{jj}$$

A variância verdadeira define a precisão e a variância comum a validade. Nesta divisão da variância pode-se ver a diferença que existe entre o conceito de validade de um teste e o de precisão. Esta é constituída pela variância verdadeira, que é composta pela variância comum e pela variância específica das variáveis do teste; ao passo que a validade é constituída exclusivamente pela variância comum, que define a covariância entre as variáveis e o fator, o que fica elucidativo se nos lembrarmos de que o fator é o traço latente do qual as variáveis observáveis são a representação comportamental. O tamanho da comunalidade (variância comum) define a qualidade da representação comportamental do traço latente pelas variáveis observáveis (itens do teste). No caso da precisão ou fidedignidade, as variáveis entram com toda a sua variância verdadeira, incluindo a específica (fica fora somente a variância erro), porque toda ela contribui para a estabilidade dos resultados (escores) do teste já que, ao

ser o teste aplicado mais de uma vez, todos os mesmos itens entram neste cômputo da precisão, coisa que não acontece com a validade, onde o crucial consiste em que cada variável represente o mesmo e único fator, isto é, só conta a variância comum entre todas as variáveis do teste.

5 – Derivação das variáveis empíricas a partir dos fatores

Segundo o modelo fatorial, conhecendo-se as cargas fatoriais (covariâncias, correlações, betas) das variáveis observáveis, podemos reproduzir as variâncias destas variáveis, bem como as intercorrelações entre elas. Um exemplo final ilustrará esta técnica. Cf. figura 11-4.

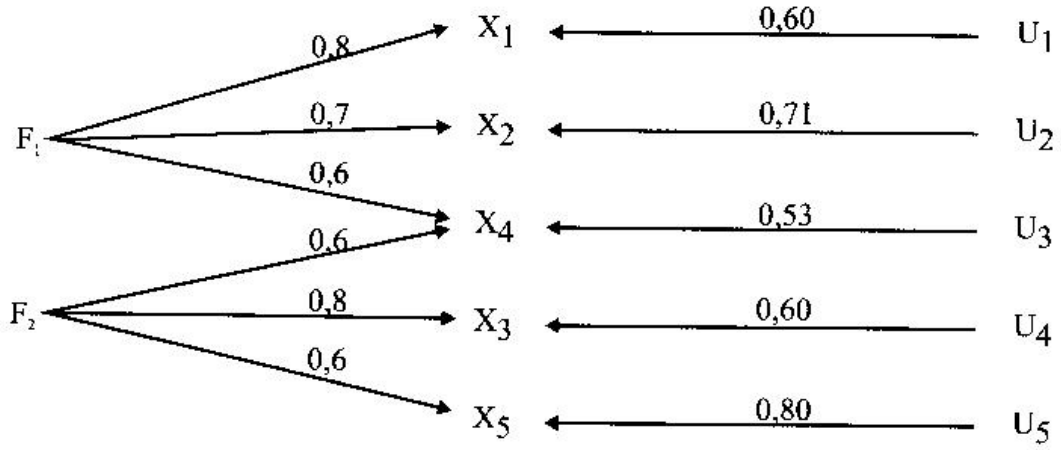


Figura 11-4. Cargas fatoriais de 5 variáveis em 2 fatores comuns

A partir desta figura, podemos construir a matriz das cargas fatoriais e a das intercorrelações entre as variáveis observáveis (Tabela 11-1 e Tabela 11-2).

Tabela 11-1. Matriz Fatorial (Matriz F)

	Fator Comum			
Var.	F_1	F_2	h^2	u^2
X_1	0,8	-	0.64	0,36
X_2	0,7	-	0.49	0,51
X_3	0,6	0,6	0.72	0,28
X_4	-	0,8	0.64	0.36
X_5	-	0,6	0.36	0,64

Tabela 11-2. Matriz das intercorrelações (Matriz R)

Var	X_1	X_2	X_3	X_4	X_5
X_1	1,00				
X_2	0,56	1,00			
X_3	0,48	0,42	1,00		
X_4	0,00	0,00	0,48	1,00	
X_5	0,00	0,00	0,36	0,48	1,00

As cargas fatoriais das variáveis X nos fatores F_1 e F_2 da tabela 11-1 são dadas diretamente pela figura 11-4. A comunalidade (h^2) da variável é a soma do quadrado das covariâncias dela com os fatores comuns. Por exemplo, no caso de X_3, a $h^2 = 0{,}6^2 + 0{,}6^2 = 0{,}72$. A unicidade ($u^2$) da variável é o complementar da comunalidade, uma vez que a variância total da variável é igual a 1, então $u^2 = 1 - h^2$, que no caso de X_3 será $1 - 0{,}72 = 0{,}28$.

Similarmente, as correlações da tabela 11-2 são obtidas através da equação 11-15, a saber, a correlação entre duas variáveis é a soma dos produtos das cargas fatoriais delas em cada fator comum. Por exemplo, a correlação entre X_2 e X_3 é a seguinte: $r_{23} = (0{,}7 \times 0{,}6) + (0{,}0 \times 0{,}6) = 0{,}42$.

Na verdade, a matriz das intercorrelações (R) entre as variáveis surge do produto de duas matrizes, a saber, a matriz fatorial F multiplicada pela sua transposta F', que, aliás, constitui a equação clássica da análise fatorial,

$$R = FF'.$$

A transposta de uma matriz é a matriz que resulta da transformação das linhas da matriz original em colunas para a matriz transposta. A multiplicação de matrizes se faz multiplicando os termos das linhas da matriz original pelos termos das colunas da matriz transposta. Exemplo:

$$\text{Matriz A: } \begin{bmatrix} 5 & 7 \\ 3 & 4 \end{bmatrix} \qquad \text{e sua transposta A': } \begin{bmatrix} 5 & 3 \\ 7 & 4 \end{bmatrix}$$

Multiplicando AA' vai dar:

$$\begin{bmatrix} 5 & 7 \\ 3 & 4 \end{bmatrix} \times \begin{bmatrix} 5 & 3 \\ 7 & 4 \end{bmatrix} = \begin{bmatrix} 5x5+7x7 & 5x3+7x4 \\ 3x5+4x7 & 3x3+4x4 \end{bmatrix} = \begin{bmatrix} 74 & 43 \\ 43 & 25 \end{bmatrix}$$

Assim, o nosso exemplo será

$$\begin{bmatrix} 0{,}8 & 0{,}0 \\ 0{,}7 & 0{,}0 \\ 0{,}6 & 0{,}6 \\ 0{,}0 & 0{,}8 \\ 0{,}0 & 0{,}6 \end{bmatrix} \times \begin{bmatrix} 0{,}8 & 0{,}7 & 0{,}6 & 0{,}0 & 0{,}0 \\ 0{,}0 & 0{,}0 & 0{,}6 & 0{,}8 & 0{,}6 \end{bmatrix} = \begin{bmatrix} 0{,}64 & 0{,}56 & 0{,}48 & 0{,}00 & 0{,}00 \\ 0{,}56 & 0{,}49 & 0{,}42 & 0{,}00 & 0{,}00 \\ 0{,}48 & 0{,}42 & 0{,}72 & 0{,}48 & 0{,}36 \\ 0{,}00 & 0{,}00 & 0{,}48 & 0{,}64 & 0{,}48 \\ 0{,}00 & 0{,}00 & 0{,}36 & 0{,}48 & 0{,}36 \end{bmatrix}$$

F F' R

A multiplicação de uma matriz pela sua transposta vai dar uma matriz quadrangular, como é o caso da matriz das intercorrelações entre os itens de um teste.

O exposto acima deve ser suficiente para entender o papel da análise fatorial no contexto da Psicometria.

APÊNDICE A
Demonstração de algumas fórmulas

Introdução

Para familiarizar o leitor, com dificuldades na mecânica da matemática, seguem as seguintes explicações sobre como deduzir simples fórmulas matemáticas. Para simplificar este trabalho de dedução, os matemáticos utilizam o conceito de *esperança matemática* ou *expectância*, abreviado como E.

A expectância de uma constante é a própria constante: E(a) = a. A expectância de uma variável é a média da distribuição desta variável. Por exemplo, a expectância da variável X é a sua média, isto é,

$$E(X) = \overline{X}.$$

Tal expressão significa que, se tivermos um número infinito de medidas do X, o valor mais representativo de todas estas medidas será o $\overline{X}$; este é o valor que resulta como a esperança matemática de todas estas medidas.

Assim, temos as seguintes equivalências:

$E(a) = a$

$E(X) = \overline{X}$

$E(aX) = aE(X)$

$E(a + X) = a + E(X)$.

Além disso, é preciso recordar que qualquer distribuição de dados de uma variável (X) pode ser exaustivamente explicada pela média e pela variância; e a relação entre duas distribuições (X e Y) pode ser dada pela correlação.

Assim, temos

Média de $X = \overline{X} = \frac{\sum X}{N} = E(X)$.

Variância de $X = s_X^2 = \frac{\sum (X - \overline{X})^2}{N} = E(X^2)$ (a variância são os desvios quadráticos médios; a saber, os desvios dos X com relação à sua média).

Correlação entre X e Y é

$r_{XY} = \frac{\sum xy}{N s_X s_Y}$, mas como

$\frac{\sum xy}{N} = Cov(X, Y)$, segue que

$r_{XY} = \frac{Cov(X, Y)}{s_X s_Y}$.

A *covariância* de X e Y é

$$Cov(X,Y) = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{N} = \frac{\sum XY}{N} - \overline{XY} = E(XY).$$

Resumindo, temos

$E(X)$ = média de X

$E(X^2)$ = variância de X

$E(XY)$ = covariância de X

$r_{XY} = \frac{Cov(X, Y)}{s_X s_Y}$, isto é, a covariância pela variância.

Além disso, nós podemos trabalhar com os dados expressos em escores padrões em lugar dos escores brutos, sem com isso perder qualquer informação dos dados. Este estratagema se torna uma grande conveniência ao se deduzirem fórmulas matemáticas, porque sabemos que em escores

padrões a média de uma distribuição é 0 e a variância é 1. Assim, as fórmulas acima podem ser expressas como

$E(X)$ = Média de $X = \overline{X} = 0$

$E(X^2)$ = Variância de $X = s_X^2 = 1$.

Dedução de Fórmulas

4.1 – E = T – V

Pelo modelo, $T = V + E$

então, $E = T - V$.

4.2 – E(E) = 0

Se (segundo 4.1) $E = T - V$

então $E(E) = E(T-V)$, isto é,

$E(E) = E(T) - E(V)$

Mas, o modelo diz que $E(T) = V$

Logo, $E(E) = V - E(V) = V - V = 0$.

4.3 – $\overline{T} = \overline{V}$

pelo modelo $T = V + E$, ou seja,

$E(T) = E(V + E) = E(V) + E(E)$

mas (segundo 4.2) $E(E) = 0$

logo $E(T) = E(V)$, isto é, $\overline{T} = \overline{V}$

4.4 – Cov (V,E) = 0

$Cov(V,E) = r_{VE} s_V s_E$ uma vez que a correlação é a covariância pela variância, isto é, $r_{VE} = \dfrac{Cov(V,E)}{s_V s_E}$

mas, segundo postulado 2 do modelo, $r_{VE} = 0$

logo, $Cov(V,E) = 0 s_V s_E = 0$.

4.5 – Cov(T,V) = Var(V)

$Cov(T,V) = E[(T - \overline{T})(V - \overline{V})]$, isto é, o produto das diferenças do T e do V com suas médias.

Resolvendo, dá

$Cov(T,V) = E(TV - T\overline{V} - \overline{T}V + \overline{TV})$. Mas,

$T = V + E$. Então,

$Cov(T,V) = E[(V+E)V - (V+E)\overline{V} - (\overline{V} + \overline{E})V + (\overline{V} + \overline{E})\overline{V}]$.

Sabendo que $\overline{E} = 0$ e, em escores padrões, também $\overline{V} = 0$,

segue que os três últimos termos desaparecem, ficando

$Cov(T,V) = E[(V^2 + VE)]$, que evidenciando dá

$Cov(T,V) = E(V^2) + E(VE)$.

Mas, E(VE) é a covariância entre V e E, a qual é 0, resulta que,

$Cov(T,V) = E(V^2) = Var(V)$.

Ou, em termos de escores brutos, mostra-se que, como

$$Cov(X,Y) = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{N} = \frac{\sum XY}{N} - \overline{XY} = E(XY) - E(X)E(Y)$$

que, em termos do modelo psicométrico, fica

$$Cov(T,V) = \frac{\sum (T - \overline{T})(V - \overline{V})}{N} = \frac{\sum TV}{N} - \overline{TV} = E(TV) - E(T)E(V)$$

e substituindo T por seu valor do modelo $T = V + E$, temos que,

$Cov(T,V) = E[(V + E)V] - E(V + E)\,E(V)$

$= E(V^2 + VE) - E(V)E(V) + E(E)E(V)$

$= E(V^2) + E(VE) - [E(V)]^2 + E(VE)$.

Como E(VE) é a covariância entre V e E, a qual é 0, segue que,

$Cov(T,V) = E(V^2) - [E(V)]^2 = Var(V)$.

4.6 – $Cov(T_i,T_j) = Cov(V_i,V_j)$

$Cov(T_i,T_j) = E(T_j) - E(T_i)E(T_j)$

substituindo T_i e T_j por valores do modelo

$Cov(T_i,T_j) = E[(V_i+E_i)(V_j+E_j)] - E(V_i + E_i)E(V_j+E_j) =$

$= E(V_iV_j + V_iE_j + V_jE_i + E_iE_j) - E(V_i)E(V_j) - E(V_i)E(E_j) - E(V_j)E(E_i) - E(E_i)E(E_j).$

Mas, pelos postulados 2 e 3 do modelo psicométrico,

$$E(V_iE_j) - E(V_i)E(E_j) = Cov(V_i,E_j) = 0$$

$$E(V_jE_i) - E(V_j)E(E_i) = Cov(V_j,V_i) = 0$$

$$E(E_iE_j) - E(E_i)E(E_j) = Cov(E_i,E_j) = 0$$

Segue que

$Cov(T_i,T_j) = E(V_iV_j) - E(V_i)E(V_j) = Cov(V_iV_j)$

Obs.: Quando se trata de formas paralelas, onde $V_i = V_j$

$Cov(T_i,T_j) = Cov(V_i,V_j) = Var(V).$

4.7 – Var(T) = Var(V) + Var(E)

Sendo $T = V + E$, sua variância será
$Var(T) = Var(V) + Var(E) + 2Cov(V,E).$
Mas, segundo 4.4, $Cov(V,E) = 0$. Logo,
$Var(T) = Var(V) + Var(E)$, isto é,

$$s_T^2 = s_V^2 + s_E^2.$$

4.8 – $r_{TE} = \frac{s_E}{s_T}$

$$r_{TE} = \frac{Cov(T,E)}{s_Ts_E} = \frac{E(TE) - E(T)E(E)}{s_Ts_E} = \frac{E[(V+E)E] - E(V+E)E(E)}{s_Ts_E}$$

$$r_{TE} = \frac{Cov(T,E)}{s_Ts_E} = \frac{E(TE) - E(T)E(E)}{s_Ts_E} = \frac{E[(V+E)E] - E(V+E)E(E)}{s_Ts_E} =$$

$$\frac{E(VE) + E(E^2) - E(V)E(E) - [E(E)]^2}{s_Ts_E}$$

Mas, $E(VE) - E(V)E(E) = Cov(V,E) = 0$ e

$E(E^2) - [E(E)]^2 = s_E^2$. Substituindo, temos

$$r_{TE} = \frac{s_E^2}{s_T s_E} = \frac{s_E}{s_T}.$$

7.1 – $r_{tt} = \dfrac{s_V^2}{s_T^2}$

Sabemos que $r_{tt} = \dfrac{Cov(T_1, T_2)}{s_{T_1} s_{T_2}}$. Mas, sendo

$Cov(T_1, T_2) = s_V^2$ e $s_{T_1} = s_{T_2}$, segue que,

$s_{T_1} s_{T_2} = s_T^2$. Então,

$$r_{tt} = \frac{s_V^2}{s_T^2}.$$

7.2 – $r_{tt} = 1 - \dfrac{s_E^2}{s_T^2}$.

De 7.1 sabemos que $r_{tt} = \dfrac{s_V^2}{s_T^2}$. Mas,

$s_V^2 = s_T^2 - s_E^2$. Então,

$$r_{tt} = \frac{s_T^2 - s_E^2}{s_T^2} = \frac{s_T^2}{s_T^2} - \frac{s_E^2}{s_T^2}$$

Dividindo por s_T^2 dá a fórmula.

7.3 – $r_{TV} = \sqrt{r_{tt}} = \dfrac{s_V}{s_T}$.

Sabemos que $r_{TV} = \dfrac{Cov(T, V)}{s_T s_V}$. Mas, segundo 4.5,

$Cov(T,V) = s_V^2$

Então, $r_{TV} = \dfrac{s_V^2}{s_T s_V}$. Dividindo por s_V dá:

$$r_{TV} = \frac{s_V}{s_T} = \sqrt{\frac{s_V^2}{s_T^2}} = \sqrt{r_{tt}} \ .$$

7.4 – EPM = $s_T\sqrt{1-r_{tt}}$ **.**

De (7.2) temos que $r_{tt} = 1 - \dfrac{s_E^2}{s_T^2} = \dfrac{s_T^2 - s_E^2}{s_T^2}$; $r_{TT}s_T^2 = s_T^2 - s_E^2$.

Resolvendo para s_E^2 dá $s_E^2 = s_T^2 - s_T^2 r_{TT} = s_T^2(1-r_{TT})$.
Assim,

$$s_E = s_T\sqrt{1-r_{tt}} \ .$$

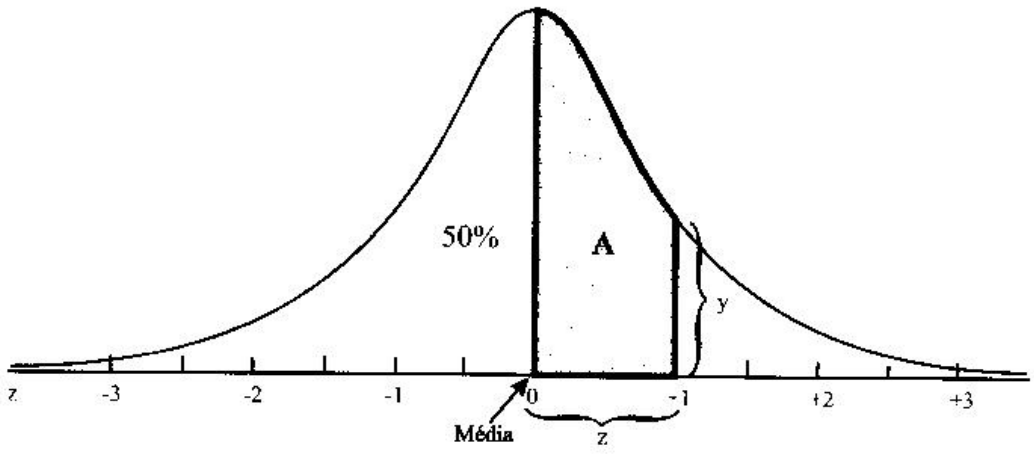
APÊNDICE B
Tabelas estatísticas

Tabela A – Distribuição normal

Na curva normal sabemos que o desvio padrão (o escore z expresso na linha de base da curva) é igual a 1 e também que a área total sob a curva é igual a 1. Há, além disso, correspondências matemáticas precisas entre as distâncias z e as áreas sob a curva. De sorte que, sabendo-se o z, pode-se conhecer diretamente o tamanho da área (o A na figura), e sabendo-se o tamanho da área (as percentagens), pode-se conhecer diretamente as distâncias z. Veja a figura a seguir: A média da distribuição tem o valor $z = 0$; abaixo da média está 50% da área e acima dela os restantes 50%; os z são referidos sempre a esta média, de sorte que z positivos indicam 50% da área + o percentual definido pelo tamanho do z (na figura, o z = 1 dá uma área extra de 34,13%, sendo o total da área assim definida de 50 + 34,13 = 84,13%) e z negativos definem áreas menores do que 50%.

Como curiosidade, veja a fórmula que define a área sob a curva:

$$\frac{1}{2}\alpha = \frac{1}{\sqrt{2\pi s}} \int_{0}^{x/s} e^{-1/2(x/s)^2} dx \text{ , donde:}$$



z Escore-padrão	A Área da média até z	y Ordenada
0,00	0,0000	0,3989
0,01	0,0040	0,3989
0,02	0,0080	0,3989
0,03	0,0120	0,3988
0,04	0,0160	0,3986
0,05	0,0199	0,3984
0,06	0,0239	0,3982
0,07	0,0279	0,3980
0,08	0,0319	0,3977
0,09	0,0359	0,3973
0,10	0,0398	0,3970
0,11	0,0438	0,3965
0,12	0,0478	0,3961
0,13	0,0517	0,3956
0,14	0,0557	0,3951
0,15	0,0596	0,3945
0,16	0,0636	0,3939
0,17	0,0675	0,3932
0,18	0,0714	0,3925
0,19	0,0753	0,3918
0,20	0,0793	0,3910
0,21	0,0832	0,3902
0,22	0,0871	0,3894
0,23	0,0910	0,3885

z Escore-padrão	A Área da média até z	y Ordenada
0,24	0,0948	0,3876
0,25	0,0987	0,3867
0,26	0,1026	0,3857
0,27	0,1064	0,3847
0,28	0,1103	0,3836
0,29	0,1141	0,3825
0,30	0,1179	0,3814
0,31	0,1217	0,3802
0,32	0,1255	0,3790
0,33	0,1293	0,3778
0,34	0,1331	0,3765
0,35	0,1368	0,3752
0,36	0,1406	0,3739
0,37	0,1443	0,3725
0,38	0,1480	0,3712
0,39	0,1517	0,3697
0,40	0,1554	0,3683
0,41	0,1591	0,3668
0,42	0,1628	0,3653
0,43	0,1664	0,3637
0,44	0,1700	0,3621
0,45	0,1736	0,3605
0,46	0,1772	0,3589
0,47	0,1808	0,3572

z Escore-padrão	A Área da média até z	y Ordenada
0,48	0,1844	0,3555
0,49	0,1879	0,3538
0,50	0,1915	0,3521
0,51	0,1950	0,3503
0,52	0,1985	0,3485
0,53	0,2019	0,3467
0,54	0,2054	0,3448
0,55	0,2088	0,3429
0,56	0,2123	0,3410
0,57	0,2157	0,3391
0,58	0,2190	0,3372
0,59	0,2224	0,3352
0,60	0,2257	0,3332
0,61	0,2291	0,3312
0,62	0,2324	0,3292
0,63	0,2357	0,3271
0,64	0,2389	0,3251
0,65	0,2422	0,3230
0,66	0,2454	0,3209
0,67	0,2486	0,3187
0,68	0,2517	0,3166
0,69	0,2549	0,3144
0,70	0,2580	0,3123
0,71	0,2611	0,3101

z Escore-padrão	A Área da média até z	y Ordenada
0,72	0,2642	0,3079
0,73	0,2673	0,3056
0,74	0,2704	0,3034
0,75	0,2734	0,3011
0,76	0,2764	0,2989
0,77	0,2791	0,2966
0,78	0,2823	0,2943
0,79	0,2852	0,2920
0,80	0,2881	0,2897
0,81	0,2910	0,2874
0,82	0,2939	0,2850
0,83	0,2967	0,2827
0,84	0,2995	0,2803
0,85	0,3023	0,2780
0,86	0,3051	0,2756
0,87	0,3078	0,2732
0,88	0,3106	0,2709
0,89	0,3133	0,2685
0,90	0,3159	0,2661
0,91	0,3186	0,2637
0,92	0,3212	0,2613
0,93	0,3238	0,2589
0,94	0,3264	0,2565
0,95	0,3289	0,2541

z Escore-padrão	A Área da média até z	y Ordenada
0,96	0,3315	0,2516
0,97	0,3340	0,2492
0,98	0,3365	0,2468
0,99	0,3389	0,2444
1,00	0,3413	0,2420
1,01	0,3438	0,2396
1,02	0,3461	0,2371
1,03	0,3485	0,2347
1,04	0,3508	0,2323
1,05	0,3531	0,2299
1,06	0,3554	0,2275
1,07	0,3577	0,2251
1,08	0,3599	0,2227
1,09	0,3621	0,2203
1,10	0,3643	0,2179
1,11	0,3665	0,2155
1,12	0,3686	0,2131
1,13	0,3708	0,2107
1,14	0,3729	0,2083
1,15	0,3749	0,2059
1,16	0,3770	0,2036
1,17	0,3790	0,2012
1,18	0,3810	0,1989
1,19	0,3830	0,1965

z Escore-padrão	A Área da média até z	y Ordenada
1,20	0,3849	0,1942
1,21	0,3868	0,1919
1,22	0,3888	0,1895
1,23	0,3907	0,1872
1,24	0,3925	0,1849
1,25	0,3944	0,1826
1,26	0,3962	0,1804
1,27	0,3980	0,1781
1,28	0,3997	0,1758
1,29	0,4015	0,1736
1,30	0,4032	0,1714
1,31	0,4049	0,1691
1,32	0,4066	0,1669
1,33	0,4082	0,1647
1,34	0,4099	0,1626
1,35	0,4115	0,1604
1,36	0,4131	0,1582
1,37	0,4147	0,1561
1,38	0,4162	0,1539
1,39	0,4177	0,1518
1,40	0,4192	0,1497
1,41	0,4207	0,1476
1,42	0,4222	0,1456
1,43	0,4236	0,1435

z Escore-padrão	A Área da média até z	y Ordenada
1,44	0,4251	0,1415
1,45	0,4265	0,1394
1,46	0,4279	0,1374
1,47	0,4292	0,1354
1,48	0,4306	0,1334
1,49	0,4319	0,1315
1,50	0,4332	0,1295
1,51	0,4345	0,1276
1,52	0,4357	0,1257
1,53	0,4370	0,1238
1,54	0,4382	0,1219
1,55	0,4394	0,1200
1,56	0,4406	0,1182
1,57	0,4418	0,1163
1,58	0,4429	0,1145
1,59	0,4441	0,1127
1,60	0,4452	0,1109
1,61	0,4463	0,1092
1,62	0,4474	0,1074
1,63	0,4484	0,1057
1,64	0,4495	0,1040
1,65	0,4505	0,1023
1,66	0,4515	0,1006
1,67	0,4525	0,0989

z Escore-padrão	A Área da média até z	y Ordenada
1,68	0,4535	0,0973
1,69	0,4545	0,0957
1,70	0,4554	0,0940
1,71	0,4564	0,0925
1,72	0,4573	0,0909
1,73	0,4582	0,0893
1,74	0,4591	0,0878
1,75	0,4599	0,0863
1,76	0,4608	0,0848
1,77	0,4616	0,0833
1,78	0,4625	0,0818
1,79	0,4633	0,0804
1,80	0,4641	0,0790
1,81	0,4649	0,0775
1,82	0,4656	0,0761
1,83	0,4664	0,0748
1,84	0,4671	0,0734
1,85	0,4678	0,0721
1,86	0,4686	0,0707
1,87	0,4693	0,0694
1,88	0,4699	0,0681
1,89	0,4706	0,0669
1,90	0,4713	0,0656
1,91	0,4719	0,0644

z Escore-padrão	A Área da média até z	y Ordenada
1,92	0,4726	0,0632
1,93	0,4732	0,0620
1,94	0,4738	0,0608
1,95	0,4744	0,0596
1,96	0,4750	0,0584
1,97	0,4756	0,0573
1,98	0,4761	0,0562
1,99	0,4767	0,0551
2,00	0,4772	0,0540
2,01	0,4778	0,0529
2,02	0,4783	0,0519
2,03	0,4788	0,0508
2,04	0,4793	0,0498
2,05	0,4798	0,0488
2,06	0,4803	0,0478
2,07	0,4808	0,0468
2,08	0,4812	0,0459
2,09	0,4817	0,0446
2,10	0,4821	0,0440
2,11	0,4826	0,0431
2,12	0,4830	0,0422
2,13	0,4834	0,0413
2,14	0,4838	0,0404
2,15	0,4842	0,0396

z Escore-padrão	A Área da média até z	y Ordenada
2,16	0,4846	0,0387
2,17	0,4850	0,0379
2,18	0,4854	0,0371
2,19	0,4857	0,0363
2,20	0,4861	0,0355
2,21	0,4864	0,0347
2,22	0,4868	0,0339
2,23	0,4871	0,0332
2,24	0,4875	0,0325
2,25	0,4878	0,0317
2,26	0,4881	0,0310
2,27	0,4884	0,0303
2,28	0,4887	0,0297
2,29	0,4890	0,0290
2,30	0,4893	0,0283
2,31	0,4896	0,0277
2,32	0,4898	0,0270
2,33	0,4901	0,0264
2,34	0,4904	0,0258
2,35	0,4906	0,0252
2,36	0,4909	0,0246
2,37	0,4911	0,0241
2,38	0,4916	0,0235
2,39	0,4916	0,0229

z Escore-padrão	A Área da média até z	y Ordenada
2,40	0,4918	0,0224
2,41	0,4920	0,0219
2,42	0,4922	0,0213
2,43	0,4925	0,0208
2,44	0,4927	0,0203
2,45	0,4929	0,0198
2,46	0,4931	0,0194
2,47	0,4932	0,0189
2,48	0,4934	0,0184
2,49	0,4936	0,0180
2,50	0,4938	0,0175
2,51	0,4940	0,0171
2,52	0,4941	0,0167
2,53	0,4943	0,0163
2,54	0,4945	0,0158
2,55	0,4946	0,0154
2,56	0,4948	0,0151
2,57	0,4949	0,0147
2,58	0,4951	0,0143
2,59	0,4952	0,0139
2,60	0,4953	0,0136
2,61	0,4955	0,0132
2,62	0,4956	0,0129
2,63	0,4957	0,0126

z Escore-padrão	A Área da média até z	y Ordenada
2,64	0,4959	0,0122
2,65	0,4960	0,0119
2,66	0,4961	0,0116
2,67	0,4962	0,0113
2,68	0,4963	0,0110
2,69	0,4964	0,0107
2,70	0,4965	0,0104
2,71	0,4966	0,0101
2,72	0,4967	0,0099
2,73	0,4968	0,0096
2,74	0,4969	0,0093
2,75	0,4970	0,0091
2,76	0,4971	0,0088
2,77	0,4972	0,0086
2,78	0,4973	0,0084
2,79	0,4974	0,0081
2,80	0,4974	0,0079
2,81	0,4975	0,0077
2,82	0,4976	0,0075
2,83	0,4977	0,0073
2,84	0,4977	0,0071
2,85	0,4978	0,0069
2,86	0,4979	0,0067
2,87	0,4979	0,0065

z Escore-padrão	A Área da média até z	y Ordenada
2,88	0,4980	0,0063
2,89	0,4981	0,0061
2,90	0,4981	0,0060
2,91	0,4982	0,0058
2,92	0,4982	0,0056
2,93	0,4983	0,0055
2,94	0,4984	0,0053
2,95	0,4984	0,0051
2,96	0,4985	0,0050
2,97	0,4985	0,0048
2,98	0,4986	0,0047
2,99	0,4986	0,0046
3,00	0,4987	0,0044
3,01	0,4987	0,0043
3,02	0,4987	0,0042
3,03	0,4988	0,0040
3,04	0,4988	0,0039
3,05	0,4989	0,0038
3,06	0,4989	0,0037
3,07	0,4989	0,0036
3,08	0,4990	0,0035
3,09	0,4990	0,0034
3,10	0,4990	0,0033
3,11	0,4991	0,0032
3,12	0,4991	0,0031
3,13	0,4991	0,0030
3,14	0,4992	0,0029
3,15	0,4992	0,0028
3,16	0,4992	0,0027
3,17	0,4992	0,0026
3,18	0,4993	0,0025
3,19	0,4993	0,0025
3,20	0,4993	0,0024
3,21	0,4993	0,0023
3,22	0,4994	0,0022
3,23	0,4994	0,0022
3,24	0,4994	0,0021
3,30	0,4995	0,0017
3,40	0,4997	0,0012
3,50	0,4998	0,0009
3,60	0,4998	0,0006
3,70	0,4999	0,0004
∞	0,5000	0,0000

Tabela B – Teste *t* e correlação

Significância a 0,05 (primeira linha) e 0,01 (segunda linha, em negrito) para vários graus de liberdade.

Graus de liber-dade	Número de variáveis									t
	2	3	4	5	6	7	9	13	25	
1	0,997	0,999	0,999	0,999	1,000	1,000	1,000	1,000	1,000	12,706
	1,000	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**63,657**
2	0,950	0,975	0,983	0,987	0,990	0,992	0,994	0,996	0,998	4,303
	0,990	**0,995**	**0,997**	**0,998**	**0,998**	**0,998**	**0,999**	**0,999**	**1,000**	**9,925**
3	0,878	0,930	0,950	0,961	0,968	0,973	0,979	0,986	0,993	3,182
	0,959	**0,976**	**0,983**	**0,987**	**0,990**	**0,991**	**0,993**	**0,995**	**0,998**	**5,841**
4	0,811	0,881	0,912	0,930	0,942	0,950	0,961	0,973	0,986	2,776
	0,917	**0,949**	**0,962**	**0,970**	**0,975**	**0,979**	**0,984**	**0,989**	**0,994**	**4,604**
5	0,754	0,836	0,874	0,898	0,914	0,925	0,941	0,958	0,978	2,571
	0,874	**0,917**	**0,937**	**0,949**	**0,957**	**0,963**	**0,971**	**0,980**	**0,989**	**4,032**
6	0,707	0,795	0,839	0,867	0,886	0,900	0,920	0,943	0,969	2,447
	0,834	**0,886**	**0,911**	**0,927**	**0,938**	**0,946**	**0,957**	**0,969**	**0,983**	**3,707**
7	0,666	0,758	0,807	0,838	0,860	0,876	0,900	0,927	0,960	2,365
	0,798	**0,855**	**0,885**	**0,904**	**0,918**	**0,928**	**0,942**	**0,958**	**0,977**	**3,499**
8	0,632	0,726	0,777	0,811	0,835	0,854	0,880	0,912	0,950	2,306
	0,765	**0,827**	**0,860**	**0,882**	**0,898**	**0,909**	**0,926**	**0,946**	**0,970**	**3,355**
9	0,602	0,697	0,750	0,786	0,812	0,832	0,861	0,897	0,941	2,262
	0,735	**0,800**	**0,836**	**0,861**	**0,878**	**0,891**	**0,911**	**0,934**	**0,963**	**3,250**
10	0,576	0,671	0,726	0,763	0,790	0,812	0,843	0,882	0,932	2,228
	0,708	**0,775**	**0,814**	**0,840**	**0,859**	**0,874**	**0,895**	**0,922**	**0,955**	**3,169**
11	0,553	0,648	0,703	0,741	0,770	0,792	0,826	0,868	0,922	2,201
	0,684	**0,753**	**0,793**	**0,821**	**0,841**	**0,857**	**0,880**	**0,910**	**0,948**	**3,106**
12	0,532	0,627	0,683	0,722	0,751	0,774	0,809	0,854	0,913	2,179
	0,661	**0,732**	**0,773**	**0,802**	**0,824**	**0,841**	**0,866**	**0,898**	**0,940**	**3,055**

13	0,514	0,608	0,664	0,703	0,733	0,757	0,794	0,840	0,904	2,160
	0,641	**0,712**	**0,755**	**0,785**	**0,807**	**0,825**	**0,852**	**0,886**	**0,932**	**3,012**
14	0,497	0,590	0,646	0,686	0,717	0,741	0,779	0,828	0,895	2,145
	0,623	**0,694**	**0,737**	**0,768**	**0,792**	**0,810**	**0,838**	**0,875**	**0,924**	**2,977**
15	0,482	0,574	0,630	0,670	0,701	0,726	0,765	0,815	0,886	2,131
	0,606	**0,677**	**0,721**	**0,752**	**0,776**	**0,796**	**0,825**	**0,864**	**0,917**	**2,947**
16	0,468	0,559	0,615	0,655	0,686	0,712	0,751	0,803	0,878	2,120
	0,590	**0,662**	**0,706**	**0,738**	**0,762**	**0,782**	**0,813**	**0,853**	**0,909**	**2,921**
17	0,456	0,545	0,601	0,641	0,673	0,698	0,738	0,792	0,869	2,110
	0,575	**0,647**	**0,691**	**0,724**	**0,749**	**0,769**	**0,800**	**0,842**	**0,902**	**2,898**
18	0,444	0,532	0,587	0,628	0,660	0,686	0,726	0,781	0,861	2,101
	0,561	**0,633**	**0,678**	**0,710**	**0,736**	**0,756**	**0,789**	**0,832**	**0,894**	**2,878**
19	0,433	0,520	0,575	0,615	0,647	0,674	0,714	0,770	0,853	2,093
	0,549	**0,620**	**0,665**	**0,698**	**0,723**	**0,744**	**0,778**	**0,822**	**0,887**	**2,861**
20	0,423	0,509	0,563	0,604	0,636	0,662	0,703	0,760	0,845	2,086
	0,537	**0,608**	**0,652**	**0,685**	**0,712**	**0,733**	**0,767**	**0,812**	**0,883**	**2,845**
21	0,413	0,498	0,552	0,592	0,624	0,651	0,693	0,750	0,837	2,080
	0,526	**0,596**	**0,641**	**0,674**	**0,700**	**0,722**	**0,756**	**0,803**	**0,873**	**2,831**
22	0,404	0,488	0,542	0,582	0,614	0,640	0,682	0,740	0,830	2,074
	0,515	**0,585**	**0,630**	**0,663**	**0,690**	**0,712**	**0,746**	**0,794**	**0,866**	**2,819**
23	0,396	0,479	0,532	0,572	0,604	0,630	0,673	0,731	0,823	2,069
	0,505	**0,574**	**0,619**	**0,652**	**0,679**	**0,701**	**0,736**	**0,785**	**0,859**	**2,807**
24	0,388	0,470	0,523	0,562	0,594	0,621	0,663	0,722	0,815	2,064
	0,496	**0,565**	**0,609**	**0,642**	**0,669**	**0,692**	**0,727**	**0,776**	**0,852**	**2,797**
25	0,381	0,432	0,514	0,553	0,585	0,612	0,654	0,714	0,808	2,060
	0,487	**0,555**	**0,600**	**0,633**	**0,660**	**0,682**	**0,718**	**0,758**	**0,846**	**2,787**

Graus de liber-dade	Número de variáveis									t
	2	3	4	5	6	7	9	13	25	
26	0,374	0,454	0,506	0,545	0,576	0,603	0,645	0,706	0,802	2,056
	0,478	**0,546**	**0,590**	**0,624**	**0,651**	**0,673**	**0,709**	**0,760**	**0,839**	**2,779**
27	0,367	0,446	0,498	0,536	0,568	0,594	0,637	0,698	0,795	2,052
	0,470	**0,538**	**0,582**	**0,615**	**0,642**	**0,664**	**0,701**	**0,752**	**0,833**	**2,771**
28	0,361	0,439	0,490	0,529	0,560	0,586	0,629	0,690	0,788	2,048
	0,463	**0,530**	**0,573**	**0,606**	**0,634**	**0,656**	**0,692**	**0,744**	**0,827**	**2,763**
29	0,355	0,432	0,482	0,521	0,552	0,579	0,621	0,682	0,782	2,045
	0,456	**0,522**	**0,565**	**0,598**	**0,625**	**0,648**	**0,685**	**0,737**	**0,821**	**2,756**
30	0,349	0,426	0,476	0,514	0,545	0,571	0,614	0,675	0,776	2,042
	0,449	**0,514**	**0,558**	**0,591**	**0,618**	**0,640**	**0,677**	**0,729**	**0,815**	**2,750**
35	0,325	0,397	0,445	0,482	0,512	0,538	0,580	0,642	0,746	2,030
	0,418	**0,481**	**0,523**	**0,556**	**0,582**	**0,605**	**0,642**	**0,696**	**0,786**	**2,724**
40	0,304	0,373	0,419	0,455	0,484	0,509	0,551	0,613	0,720	2,021
	0,393	**0,454**	**0,494**	**0,526**	**0,552**	**0,575**	**0,612**	**0,667**	**0,761**	**2,704**
45	0,288	0,353	0,397	0,432	0,460	0,485	0,526	0,587	0,696	2,014
	0,372	**0,430**	**0,470**	**0,501**	**0,527**	**0,549**	**0,586**	**0,640**	**0,737**	**2,690**
50	0,273	0,336	0,379	0,412	0,440	0,464	0,504	0,565	0,674	2,008
	0,354	**0,410**	**0,449**	**0,479**	**0,504**	**0,526**	**0,562**	**0,617**	**0,715**	**2,678**
60	0,250	0,308	0,348	0,380	0,406	0,429	0,467	0,526	0,636	2,000
	0,325	**0,377**	**0,414**	**0,442**	**0,466**	**0,488**	**0,523**	**0,577**	**0,677**	**2,660**
70	0,233	0,286	0,324	0,354	0,379	0,401	0,438	0,495	0,604	1,994
	0,302	**0,351**	**0,386**	**0,413**	**0,436**	**0,456**	**0,491**	**0,544**	**0,644**	**2,648**
80	0,217	0,269	0,304	0,332	0,356	0,377	0,413	0,469	0,576	1,990
	0,283	**0,330**	**0,362**	**0,389**	**0,411**	**0,431**	**0,464**	**0,516**	**0,615**	**2,638**
90	0,205	0,254	0,288	0,315	0,338	0,358	0,392	0,446	0,552	1,987
	0,267	**0,312**	**0,343**	**0,368**	**0,390**	**0,409**	**0,441**	**0,492**	**0,590**	**2,632**
100	0,195	0,241	0,274	0,300	0,322	0,341	0,374	0,426	0,530	1,984

	0,254	**0,297**	**0,327**	**0,351**	**0,372**	**0,390**	**0,421**	**0,470**	**0,568**	**2,626**
125	0,174	0,216	0,246	0,269	0,290	0,307	0,338	0,387	0,485	1,979
	0,228	**0,266**	**0,294**	**0,316**	**0,335**	**0,352**	**0,381**	**0,428**	**0,521**	**2,616**
150	0,159	0,198	0,225	0,247	0,266	0,282	0,310	0,356	0,450	1,976
	0,208	**0,244**	**0,270**	**0,290**	**0,308**	**0,324**	**0,351**	**0,395**	**0,484**	**2,609**
200	0,138	0,172	0,196	0,215	0,231	0,246	0,271	0,312	0,398	1,972
	0,181	**0,212**	**0,234**	**0,253**	**0,269**	**0,283**	**0,307**	**0,347**	**0,460**	**2,601**
300	0,113	0,141	0,160	0,176	0,190	0,202	0,223	0,258	0,332	1,968
	0,148	**0,174**	**0,192**	**0,208**	**0,221**	**0,233**	**0,253**	**0,287**	**0,359**	**2,592**
400	0,098	0,122	0,139	0,153	0,165	0,176	0,194	0,225	0,291	1,966
	0,128	**0,151**	**0,167**	**0,180**	**0,192**	**0,202**	**0,220**	**0,250**	**0,315**	**2,588**
500	0,088	0,109	0,124	0,137	0,148	0,157	0,174	0,202	0,262	1,965
	0,115	**0,135**	**0,150**	**0,162**	**0,172**	**0,182**	**0,198**	**0,225**	**0,284**	**2,586**
1,000	0,062	0,077	0,088	0,097	0,105	0,112	0,124	0,144	0,188	1,962
	0,081	**0,096**	**0,106**	**0,115**	**0,122**	**0,129**	**0,141**	**0,160**	**0,204**	**2,581**
∞										1,960
										2,576

APÊNDICE C
Programas de computador para Psicometria

Os seguintes pacotes de computador você pode conseguir com

Assessment Systems Corporation
2233 University Avenue, Suite 200
St. Paul, MN, 55114-1629 USA
www.assess.com
E-mail: sales@assess.com
Fax: 651/6470412

A – Para bancos de itens, testagem em linha (computadorizada), construção de testes:

– *FastTEST Professional*: para criar bancos de itens e testagem em linha (US$995.00)

– *C-Quest*: para criar testes, provas e surveys computadorizados (US$390.00).

– *WinAsks Professional*: para criar e analisar questionários e surveys multimídia (US$995.00) (US$99.00).

– *Edwin Professional*: para criar questionários (US$355.00).

– *The Examiner*: para criar sistemas de testes computadorizados com multimídia (US$875.00).

– *Contest*: para montagem de testes (assembly testing; US$1,690.00).

– *MicroCAT* (*Computerized Testing System with Developers Licence*): Sistema computadorizado de testagem que dá suporte a todo o sistema de testagem; ele pode criar e manter banco de itens, desenvolver e criar formulário de testes, desenvolver e aplicar testes computadorizados, efetuar análises clássicas, bem como as análises logísticas da TRI. Oferece modelos para criação de testes; serve para criar banco de itens e testes computadorizados (US$1,000.00).

– *FastTEST*: para criar banco de itens e imprimir testes on-line ou papel (US$199.00).

B – Para análise clássica de itens

– *ASC Item and Test Analysis Package*: um pacote que contém os seguintes softs (US$ 950.00):

  – *ITEMAN*: analisa itens dicotômicos, de múltipla escolha e de survey.

  – *RASCAL*: analisa os parâmetros dos itens segundo o modelo de Rasch.

  – *XCALIBRE*: analisa os itens para 1-, 2-, e 3-parâmetros da TRI.

  – *ASCAL*: utiliza o modelo bayesiano modal para análise da TRI.

  – *TESTINFO (Test Information Program)*: estima a curva de informação e a fidedignidade do teste.

  – *TESTVAL (Test Validation Program)*: calcula estatísticas descritivas e de regressão bivariada e múltipla.

  – *SCOREALL (Test Scoring Program)*: calcula as aptidões dos sujeitos baseadas na TRI.

C – Para análise da Teoria de Resposta ao Item:

– *XCALIBRE*: analisa itens binários pelos modelos de 1-, 2- e 3-parâmetros da TRI (US$ 399.00).

– *Multilog (Analysis of multiple-category response data)*: analisa itens de resposta múltipla tipo Likert. Utiliza o método da máxima verossimilhança marginal na estimação dos parâmetros e a máxima verossimilhança e a estimação Bayesiana na fase de produzir os escores US$ 270.00).

– *BILOG-W* e *MG (Analysis of binary response data)*: analisa itens binários pelos modelos de 1-, 2-, e 3-parâmetros (W: US$ 400.00; MG: US$ 385.00).

R.J. Mislevy, Educational Testing Services, Princeton, USA.

R.D. Bock, University of Chicago, USA.

– *PARSCALE*: analisa itens e testes com respostas múltiplas (US$ 385.00).

– *PARELLA* (elaborado por Ivo Molenaar e outros): análise paramétrica de itens para medida de atitudes e preferências (US$ 280.00).

– *ASCAL*: (US$ 289.00).

– *MSP*: análise de TRI não paramétrica (US$ 280.00).

– *POSTSIM* (Post-hoc Adaptive Testing Simulation): para desenvolver testes adaptativos (US$ 75.00).

– *PARDSIM* (Parameter and Response Data Simulation Program): para gerar parâmetros dos itens e dos tetas via TRI, usando simulação Monte Carlo (US$ 75.00).

– *TESTINFO* (Test Information Program): analisa parâmetros de itens dicotômicos via TRI (US$ 75.00).

– *TESTVAL* (Test Validation Program): faz análise de validade de critério para testes aplicados (US$ 75.00).

– *SCOREALL* (The Scoring Program): calcula escores de testes aplicados, utilizando TRI com máxima verossimilhança e Bayes (US$ 99.00).

D – Para avaliar a dimensionalidade da estrutura latente:

– *DIMTEST*: verifica a unidimensionalidade de testes com itens dicotômicos (US$ 150.00).

– *POLY-DIMTEST*: verifica a unidimensionalidade para testes com itens politômicos (US$ 150.00).

– *CONCOV*: análise não paramétrica da unidimensionalidade de um teste com itens dicotômicos (US$ 100.00).

– *HCA/CCPROX*: análise não paramétrica da dimensionalidade de um teste com itens dicotômicos (US$ 125.00).

– *DETECT*: analisa um teste de múltiplos conteúdos em agrupamentos homogêneos (US$ 150.00).

E – Para avaliar a função diferencial dos itens (DIF):

– *DIFCOMP*: analisa o DIF para o modelo TRI de 2-parâmetros (US$ 60.00).

– *SIBTEST*: análise não paramétrica do DIF (US$ 195.00).

– *Polytomous SIBTEST*: o SIBTEST para itens politômicos (US$ 195.00).

F – Para efetuar análise fatorial e dos componentes:

– *MicroFact*: faz análise fatorial para itens dicotômicos e politômicos (US$ 299.00).

– *TestFACT*: faz análise fatorial para itens dicotômicos e a full-information TRI (US$ 385.00).

– *SCA*: faz análise dos componentes para dois ou mais grupos que têm o mesmo sentido (US$ 280.00).

OUTROS SOFTS:

PC-BILOG: R. Mislevy & R.D. Bock (1986). Mooresville, In: Scientific Software, Inc.

RUMM for Windows: analisa itens politômicos pelo modelo de Rasch. D. Andrich, B. Sheridan, & G. Luo, Edith Cowan University, Pearson Street, Churchlands, Western Australia (US$ 500.00).

SPSS for Windows (*Statistical Package for the Social Sciences*).

Pacote estatístico para Ciências Sociais, que oferece uma extensa variedade de ferramentas estatísticas, de gráficos e relatórios. Os procedimentos estatísticos variam de estatísticas paramétricas e não paramétricas simples até estatísticas avançadas como análises multivariadas, composto pelos seguintes módulos: SPSS Base, SPSS Professional Statistics, SPSS Advanced Statistics, SPSS Tables, SPSS Categories, SPSS Trends, SPSS Chaid, SPSS Exact Tests.

MUDFOLD 4.0 (*Multiple Unidimensional Unfolding*).

Programa para analisar proximidade de dados (atitudes, preferências e múltipla escolha) com base no modelo de *desdobramento* de Coombs. Próprio para ordenar categorias, escalas tipo Likert ou dados dicotômicos, encontrando o máximo de subconjuntos de estímulos que pode ser representado em uma dimensão desdobrada.

LISREL 8 / PRELIS 2.

Analisa relações estruturais, bem como outras análises multivariadas, tais como análise de percurso, análise de estruturas de médias e análise de multiamostras. PRELIS está incluído com amplo leque de opções de pré-processamento.

EQS 5.

Utiliza o modelo Bentler-Weeks para análise dos modelos das equações estruturais. Realiza regressão múltipla, regressão multivariada, análise fatorial confirmatória, análise de estrutura de médias e comparações múltiplas da população.

BIMAIN 2.

Faz análises de TRI de grupos múltiplos para 1-, 2-, 3-parâmetros e a manutenção de testes com itens binários.

CorelDraw 10.0.

Para produção de gráficos, desenhos e filmes.

Logist.

Pacote para análise TRI de 1-, 2-, 3-parâmetros para mainframe.

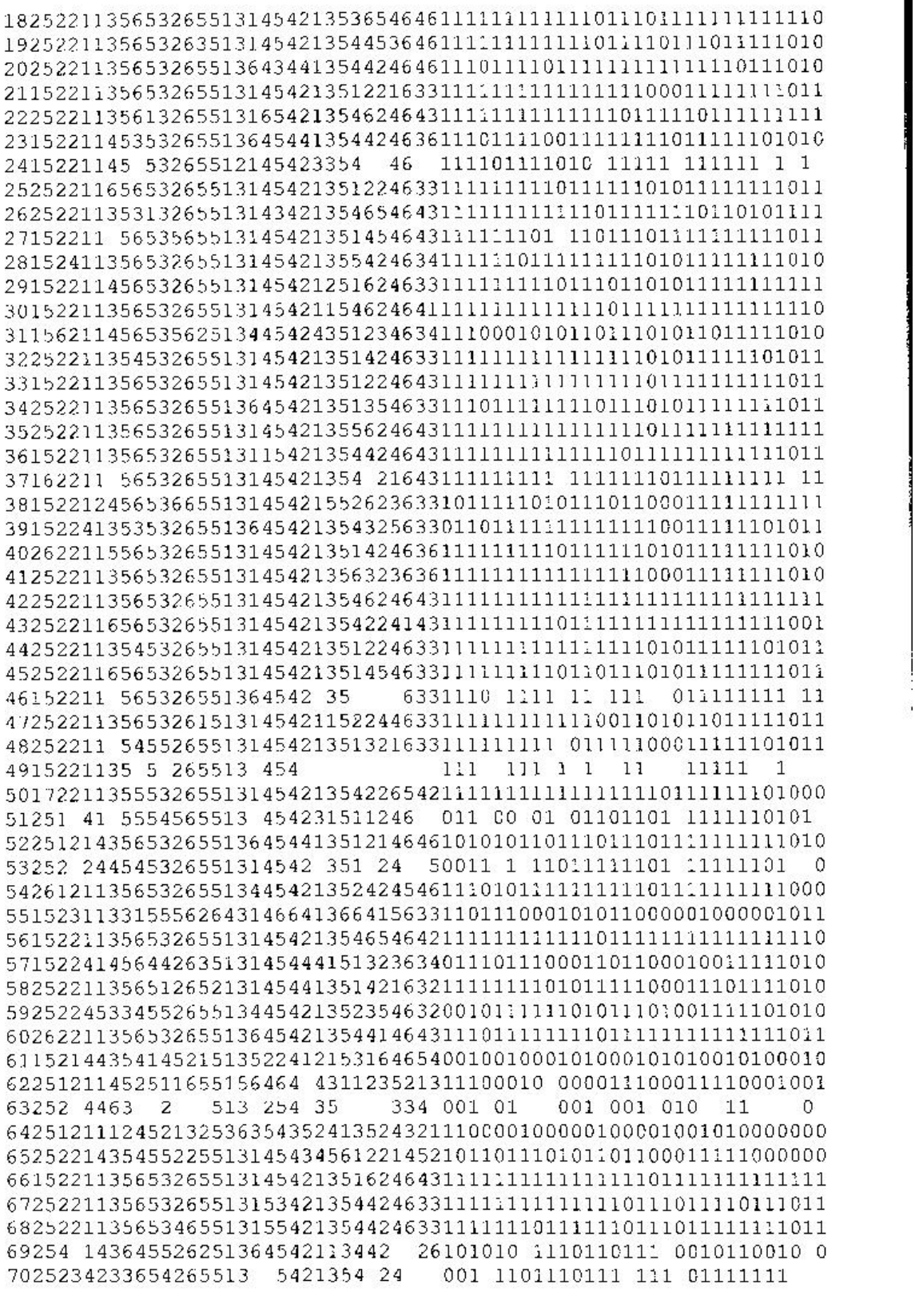
APÊNDICE D
Programas em SPSS para análise do TNVRA

I – Banco de itens do TNVRA (250casos)

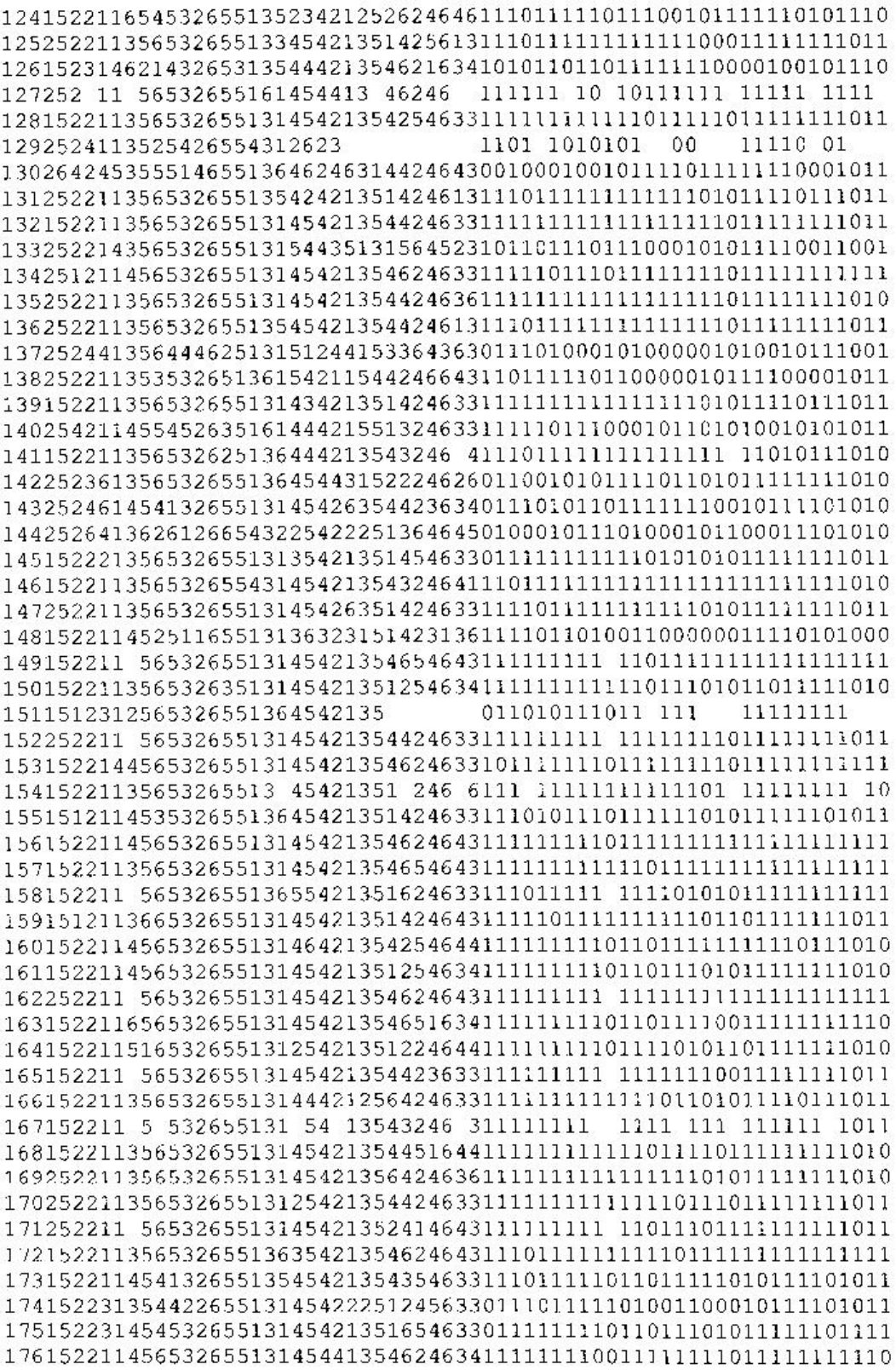
Leitura dos dados:

- colunas 1 a 4: sujeito (1 a 250)
- coluna 5: sexo do sujeito (1=masculino, 2=feminino)
- coluna 6: escolaridade (II grau e superior)
- colunas 7 a 36: respostas dos sujeitos aos 30 itens
- colunas 37 a 66: respostas corrigidas (1=acertou, 2=errou, em branco=dado omisso)

```
125221135653265513 454213516246 3111 11111111111101 11111111111
2162511356532655131454213542246361111110111111111111011111111010
3151 113563265513645421 5   46 3111010 11111  11 1 11111111 11
4152211356532655136354213544246431110111111111011111111111111011
5252211356532655131454213514246331111111111111111010111111111011
6261411356562652131454413514446331111100101010111010111011111011
7252 113565326551314542135l  46 6111111 11111 11101 11111111 10
8252211356532655131454213543216361111111111111111100111111111010
9251211356532655131454213546246361111101111111111111011111111110
1025211135 5  6455 1264631151666 41101010 01  000100 110100 1010
1125221135653265513645451351324632111011110111111101011111111010
12252211 56532655131454213516246431111111111 11111101111111111111
1315221141653265513145421354 24613111111111101111111100111111111
1425221135653265513145421354424644111111111111111111111111111010
1515221135653265513145421351354633111111111111101110101111111011
1625221145653265513145421354621643111111111101111111011111111111
1725221135653265513155421354423633111111111111111011001111111011
```



71252 1 33653265513145421351424641 1111 11111111101101111111101
721522113565326551314542135445464311111111111101111111111111011
731522114565326551314542115442464411111111101110111111111111010
741515114565326551314542135422461111111001101111111101111111010
751522113565326551314544135142163311111111011111110001111111011
762522113565326551314542125442464411111111111111011111111111010
771522113565326551314542115336463611111111111110011010111111111010
78252211551516356632146322436 6110 01100001 00110 11100000 10
791523413544352554354542123525463601001100111100110101011000010
802522113565326511314244135165464511111111101110111011111001111110
811522113535326551315444115262464611111111101111001011111101011110
821522113565326551314542135442464611111111101111001011111101011110
832522113565426551364542 33 4546341110 11111010111 10111110111010
842522133565326551332542135442463310101111111111101110111111111011
852522113565326551314542135242663611111111111111101110111111111011
861516113354626551122442422162362211100001110010010000011000011110
872546116562526551313542113446562611111001100100011001011011010
882522113545326551314242135 246 31111111111111111101 11110101 11
892521411325264363142263345412 46 30100010000000100011 01000100011
902622113565326551314544135462 4 331111111101111111111011111111111 1
912511151545316651315542215346433110110000101100010101101110100 0
92251123121545 6 563 444211541246 4000 10001001101111 01 10101010
931522114525326551314542135141463611111111101101110101111110101 0
942522113565326551314544135442463411111111101111111111011111111010
951522113565326551314542135422463311111111111111111111011111111011
962522113565326651314542133442463311111111111111111111011111111011
972522113565326551314542135142464311111111111111111101101101101 1
982522113565326551314542135134463611111111111111101111111111111011
992542114565626551334544135442453311101011000111111101111111111010
100252611346532655131454213511246431111110111111111101101111111001
101252211354132655133454213543246461110111111111111111101111111011
102152213464552655136154261514246331010011110011000101001111101011
103151211356532655131454213542216451111101111111111110111111111010
104252211456532655131454213542236351111111110111111110011111111010
105251 113565326551314542125262463311111 1111110110101111111111111
106152211356532655131354213544216331111111111111111101100111111111011
107151211353532655 13 2544132 546 6111 101101110101 1 11111001 10
108252314565326551314542135462464601111111101111111111111111111110
109152211356532655131454313516246431111111101111111101111111111111
110252211354532655131454111523566341111111101110011101111111111111
111252211456532655431454113514646331101111100110111010111111111011
112152415356532655132454213515246661010110111111111101011111111010
113252211654532645434554412216636661100111100110001000110110011110
114255111355542655131454212543643341111100111010011110111111101000
115252211356532655131454213544246331111111111111111111011111111011
116251364632422655131554233564366330011000110010101000001111010111
117252411352542655131454213546446331111110111010111111011111101111
118152511456532655131454213544246331111110110111111110111111111011
119252411356532655111454213536246331111110111011111110111111111111
120152514334562655131424253143216441011010111011111110101111110001010
121252211356532655131454213543246331111111111111111110111111111011
122162211356532655136454413516246461110111101111111110111111111110
123252211356562655131534223544166331111011111010101100111101110111



```
17725234453511366213142463352442554001101000001011100000000101000
17815241134543265513145422254423634111101011111101110000111101010
17915221145514265513145421554324631111111111001101111010111101010
180252231 56532655131454213566246430111111 11111101111111111111
18125221135353265513145421351224636111111111111111101011111101010
18215221145556265513156456652324145111101110001100101111110101000
18315221135653265513145421655624633111111111111101101011111111111
18415242555353265513245421351621143001011011011111100111111101101
18515221145653265513145421351424643111111111011111101111111111011
186153211 555326551314542135115 4   111110111 11011101 111111010
18714221465453265513145422 35 223645101101111010111111 0111111101010
18825221135653265513645421351624643111011111111111102111111111111
18916221145653265513145435356246341111011100111111010111111111110
19025221145613265513145425254221636111101111011101110010111111010
19125221145653265513145421354424634111111111011111111101111111010
19225221135453265513145421354334645111111111111011111111111101010
19325221135653265513145421354424636111111111111111111011111111010
19415221135653265513145441353244644111111110111011101111111111010
19525224135413265513145421352421145011111111111111100110111101000
19615121135653265513145421354624636111110111111111111011111111110
19715224115454266543  5421154 24633010 1111100110 1110110111101 11
19815221125153266543445413352124246110001110011111101111011101000
19915221135653265513145421351324634111111111111111010111111113010
20015221135653265513145421354624634111111111111111110111111111110
20115222135653265513145421352221633011111111111111000111111111011
20215226135653265513645421351 6 1643011011111111 1110011111111111
20315221135653262513145441354424644111111110111111111111110111110
20415221135653265113432421155224636111011111111100101011100111010
20515221145653265513145421351424646111111110111111011111111111010
20615221145653265513145421354624643111111111101111111111111111111
20715221145653265513145421351621633111111111101111110001111111111
20825221135653265513145421654324644111111111111101111111111111010
20925221165653265513145421353246441111111110111111011111111111010
21015122165653265513145421354625643011110111011111110111111111111
21115221145653265513145431354624643111111110011111311111111111111
21215222145653265513145421354414644011111111101101111111111111010
21315121135653265513145421354624644111110111111111111111111111110
21415221145653265513145421354624643111111111011111111111111111111
215151211 5653265513 4541135162 631111 10110 1111110 0111111111110
21615221145653265513145421354624643111111111101111111111111111111
217152211 56532655131454211544546331111111111 1100111101111111011
21815221145653265513145421351424643111111111011111101111111111011
21915221145653265513145421351634643111111111011011101111111111111
22015221145653265513145421354624633111111111011111111011111111111
22115221145653265513145421351624635111111111011111101011111111110
22215221165653265513145421354654636111111111011011111011111111110
22315221145653265513145421354654636111111111101101111101111111110
22415221145653265513645421354344643111101111101101111111111111011
22515221145553 551314242135412413311111111101111111101111010 001
22615254223451265423142623222423646000101011001101000101100001010
22715221145653265513145421151 24643111121111011101101111111111 11
22815221145653 655131454213546246441111111 10111211111111211111110
22915221135353265513145441354423632111111110111211110011111010010
```

```
230152211456532655131454213546246431111111101111111111111111111
231152211456532645131454213546546361111111101101111011011111110
232151321356532655136454213512246230110100111111111010111111111011
233152211456532655131454213544246331111111110111111110111111111011
234152211456532655131454213543246361111111110111111110111111111010
235152211456532655131454413543436431111111100110111101111111111011
236152211 56532655131 542135462164311111111 1111 110111111111111
237152211456532655131454213541236431111111111011111110111111111011
238152211456532655131354213546246331111111111011110111011111111111
239252211456532655131454213544246331111111111011111111011111111011
240252211456532655131454213544246431111111111011111111111111111011
241152211456532655131454213514546431111111111011011101111111111011
242252211456532655131454213544246331111111111011111111011111111011
243152211355532655131454223532236341111011111111110001111110101O
244152211456532655131154213546241 611111111011110111 11111111100
245152211354112655131344243516516331111011111010101000101101011111
246152211356532655131454213546246431111111111111111111111111111111
247152211354532655131454213566246331111111111111111110101111110111
248151211353532651431454213546246331101101111111111110111011011111
249152211356532655131454213543216331111111111111111110011111111011
250152211556532655136454223542346331110011110110111110111111111011
```

II – Análises do TNVRA (programas do SPSS)

```
COMMENT correção dos itens do teste
RECODE
 v1 v9 v16 v30 v33 v35 v38 v47 v51  (1=1)  (2 thru 6=2)
 (ELSE=SYSMIS) into vi1 vi9 vi16 vi30 vi33 vi35 vi38 vi47 vi51.
EXECUTE .
RECODE
 v6 v7 v18 v21 v29 v32 v37 v48 v53 v55  v56 (2=1)  (1,3 thru 6=2)
 (ELSE=SYSMIS) into vi6 vi7 vi18 vi21 vi29 vi32 vi37 vi44 vi48 vi53
 vi55 vi56.
EXECUTE .
RECODE
 v2 v11 v12 v17 v23 v26 v39 v45 v52 v58 (3=1)  (1,2,4 thru 6=2)
 (ELSE=SYSMIS) into vi2 vi11 vi12 vi17 vi23 vi26 vi39 vi45 vi52 vi58.
EXECUTE .
RECODE
 v5 v8 v15 v24 v27 v34 v36 v41 v46 v50 v60 (4=1)  (1 thru 3,5,6=2)
 (ELSE=SYSMIS) into vi5 vi8 vi15 vi24 vi27 vi34 vi36 vi41 vi46 vi50
 vi60.
EXECUTE .
RECODE
```

```
 v4 v14 v19 v20 v25 v40 v43 v54 v57 (5=1) (1 thru 4,6=2)
 (ELSE=SYSMIS) into vi4 vi14 vi19 vi20 vi25 vi40 vi43 vi54 vi57.
EXECUTE .
RECODE
 v3 v10 v13 v22 v28 v31 v42 v44 v49 v59 (6=1) (1 thru 5=2)
 (ELSE=SYSMIS) into vi3 vi10 vi13 vi22 vi28 vi31 vi42 vi49 vi59.
EXECUTE .

COMMENT Análise Fatorial PAF (Eixos principais) do TNVRA
FACTOR
 /VARIABLES vi3 vi4 vi14 vi16 vi21 vi1 vi2 vi10 vi26 vi5 vi9 vi15 vi20
 vi22
 vi17 vi19 vi24 vi27 vi29 vi6 vi8 vi12 vi13 vi18 vi23 vi7 vi11 vi25 vi28
 vi30
 /MISSING PAIRWISE /ANALYSIS vi3 vi4 vi14 vi16 vi21 vi1 vi2 vi10
 vi26 vi5
 vi9 vi15 vi20 vi22 vi17 vi19 vi24 vi27 vi29 vi6 vi8 vi12 vi13 vi18 vi23
 vi7
 vi11 vi25 vi28 vi30
 /PRINT UNIVARIATE INITIAL DET KMO EXTRACTION
 /FORMAT SORT
 /PLOT EIGEN
 /CRITERIA FACTORS(1) ITERATE(25)
 /EXTRACTION PAF
 /ROTATION NOROTATE
 /METHOD=CORRELATION .

COMMENT Análise da consistência do TNVRA alfa de Cronbach e
 lambda de Guttman
RELIABILITY
 /VARIABLES=vi3 vi4 vi14 vi16 vi21 vi1 vi2 vi10 vi26 vi5 vi9 vi15 vi20
 vi22
 vi17 vi19 vi24 vi27 vi29 vi6 vi8 vi12 vi13 vi18 vi23 vi7 vi11 vi25 vi28
 vi30
 /FORMAT=NOLABELS
 /SCALE(GUTTMAN)=ALL/MODEL=GUTTMAN
```

```
 /STATISTICS=DESCRIPTIVE SCALE
 /SUMMARY=TOTAL .

COMMENT Criação do escore fatorial do TNVRA
COUNT
 Escore = vi3 vi4 vi14 vi16 vi21 vi1 vi2 vi10 vi26 vi5 vi9 vi15 vi20 vi22
 vi17 vi19 vi24 vi27 vi29 vi6 vi8 vi12 vi13 vi18 vi23 vi7 vi11 vi25 vi28
  vi30
  (1) .
VARIABLE LABELS Escore 'Escore fatorial do TNVRA' .
EXECUTE .

COMMENT Criação do escore z do escore fatorial do TNVRA
DESCRIPTIVES
 VARIABLES=escore /SAVE
 /STATISTICS=MEAN STDDEV MIN MAX SEMEAN KURTOSIS
  SKEWNESS .

COMMENT Criação dos escores percentílicos do TNVRA
FREQUENCIES
 VARIABLES=escore /FORMAT=NOTABLE
 /NTILES= 100
 /ORDER= ANALYSIS .

COMMENT Criação do escore normalizado T = 50 + 10z do TNVRA
COMPUTE escoreT = 50+10 * zescore .
EXECUTE .

COMMENT Criação dos grupos-critério
COMPUTE grupo = escore .
EXECUTE .
RECODE
 grupo (Lowest thru 10=1) (20 thru Highest=2) .
EXECUTE .
```

COMMENT **Discriminação dos itens** do TNVRA por grupos-critério (teste t)
T-TEST
GROUPS=grupo(1 2)
/MISSING=ANALYSIS
/VARIABLES=vi3 vi4 vi14 vi16 vi21 vi1 vi2 vi10 vi26 vi5 vi9 vi15 vi20 vi22
vi17 vi19 vi24 vi27 vi29 vi6 vi8 vi12 vi13 vi18 vi23 vi7 vi11 vi25 vi28 vi30
/CRITERIA=CIN(.95) .

COMMENT **Dificuldade dos itens** do TNVRA (percentual de acertos)
FREQUENCIES
VARIABLES=vi3 vi4 vi14 vi16 vi21 vi1 vi2 vi10 vi26 vi5 vi9 vi15 vi20 vi22
vi17 vi19 vi24 vi27 vi29 vi6 vi8 vi12 vi13 vi18 vi23 vi7 vi11 vi25 vi28 vi30
/ORDER= ANALYSIS .

COMMENT **Análise gráfica** do TNVRA

```
COUNT
 soma = vi3 vi4 vi14 vi16 vi21 vi1 vi2 vi10 vi26 vi5 vi9 vi15
  vi20 vi22 vi17 vi19 vi24 vi27 vi29 vi6 vi8 vi12 vi13 vi18 vi23
  vi7 vi11 vi25 vi28 vi30  (1).
VARIABLE LABELS soma 'escore total'.
EXECUTE.
FORMATS soma(f2).
EXECUTE.

DO REPEAT x1=v1 to v30/y1=v1_1 to v1_30/y2=v2_1 to v2_30/y3=v3_1
  to v3_30/y4=v4_1 to v4_30/y5=v5_1 to v5_30/y6=v6_1 to v6_30.
IF (x1 eq 1)y1=1.
IF (x1 ne 1)y1=0.
IF (x1 eq 2)y2=1.
IF (x1 ne 2)y2=0.
```

```
IF (x1 eq 3)y3=1.
IF (x1 ne 3)y3=0.
IF (x1 eq 4)y4=1.
IF (x1 ne 4)y4=0.
IF (x1 eq 5)y5=1.
IF (x1 ne 5)y5=0.
IF (x1 eq 6)y6=1.
IF (x1 ne 6)y6=0.
END REPEAT.
EXECUTE.

GRAPH /LINE=MEAN(v2_1 v1_1 v3_1 v4_1 v5_1 v6_1) BY soma.
GRAPH /LINE=MEAN(v2_2 v1_2 v3_2 v4_2 v5_2 v6_2) BY soma.
GRAPH /LINE=MEAN(v1_3 v2_3 v3_3 v4_3 v5_3 v6_3) BY soma.
GRAPH /LINE=MEAN(v1_4 v2_4 v3_4 v4_4 v5_4 v6_4) BY soma.
GRAPH /LINE=MEAN(v3_5 v1_5 v2_5 v4_5 v5_5 v6_5) BY soma.
GRAPH /LINE=MEAN(v5_6 v1_6 v2_6 v3_6 v4_6 v6_6) BY soma.
GRAPH /LINE=MEAN(v6_7 v1_7 v2_7 v3_7 v4_7 v5_7) BY soma.
GRAPH /LINE=MEAN(v5_8 v1_8 v2_8 v3_8 v4_8 v6_8) BY soma.
GRAPH /LINE=MEAN(v3_9 v1_9 v2_9 v4_9 v5_9 v6_9) BY soma.
GRAPH /LINE=MEAN(v2_10 v1_10 v3_10 v4_10 v5_10 v6_10) BY soma.
GRAPH /LINE=MEAN(v6_11 v1_11 v2_11 v3_11 v4_11 v5_11) BY soma.
GRAPH /LINE=MEAN(v5_12 v1_12 v2_12 v3_12 v4_12 v6_12) BY soma.
GRAPH /LINE=MEAN(v5_13 v1_13 v2_13 v3_13 v4_13 v6_13) BY soma.
GRAPH /LINE=MEAN(v1_14 v2_14 v3_14 v4_14 v5_14 v6_14) BY soma.
GRAPH /LINE=MEAN(v3_15 v1_15 v2_15 v4_15 v5_15 v6_15) BY soma.
GRAPH /LINE=MEAN(v1_16 v2_16 v3_16 v4_16 v5_16 v6_16) BY soma.
GRAPH /LINE=MEAN(v4_17 v1_17 v2_17 v3_17 v5_17 v6_17) BY soma.
GRAPH /LINE=MEAN(v5_18 v1_18 v2_18 v3_18 v4_18 v6_18) BY soma.
GRAPH /LINE=MEAN(v4_19 v1_19 v2_19 v3_19 v5_19 v6_19) BY soma.
GRAPH /LINE=MEAN(v2_20 v1_20 v3_20 v4_20 v5_20 v6_20) BY soma.
GRAPH /LINE=MEAN(v1_21 v2_21 v3_21 v4_21 v5_21 v6_21) BY soma.
GRAPH /LINE=MEAN(v3_22 v1_22 v2_22 v4_22 v5_22 v6_22) BY soma.
GRAPH /LINE=MEAN(v5_23 v1_23 v2_23 v3_23 v4_23 v6_23) BY soma.
GRAPH /LINE=MEAN(v4_24 v1_24 v2_24 v3_24 v5_24 v6_24) BY soma.
GRAPH /LINE=MEAN(v6_25 v1_25 v2_25 v3_25 v4_25 v5_25) BY soma.
```

```
GRAPH /LINE=MEAN(v2_26 v1_26 v3_26 v4_26 v5_26 v6_26) BY soma.
GRAPH /LINE=MEAN(v4_27 v1_27 v2_27 v3_27 v5_27 v6_27) BY soma.
GRAPH /LINE=MEAN(v6_28 v1_28 v2_28 v3_28 v4_28 v5_28) BY soma.
GRAPH /LINE=MEAN(v4_29 v1_29 v2_29 v3_29 v5_29 v6_29) BY soma.
GRAPH /LINE=MEAN(v3_30 v1_30 v2_30 v4_30 v5_30 v6_30) BY soma.
```

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

Para Análise dos itens **via TRI com a Xcalibre do ASC,** o arquivo de dados deve começar com as 4 linhas abaixo assinaladas:

```
030 o n 36                          no. Itens,omit,não atingidos, pulo
22113565326551314542135462463 6      chave
66666666666666666666666666666666     no. Alternativas
YYYYYYYYYYYYYYYYYYYYYYYYYYYYYY       itens a incluir
   125221135653265513 454213516246 3111 11111111111101 11111111111
   216251135653265513145421354224636111111011111111111011111111010
   3151 1135653265513645421 5   46 3111010 11111  11 1 11111111 11
   415221135653265513635421354424643111011111111110111111111111011
  ........
```

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

Para análise da **Full Information** do TNVRA, pelo pacote estatístico TESTFACT, os comandos são os seguintes (o banco de dados é montado como no ponto I acima):

```
>TITLE
     TNVRA FULL-INFORMATION FACTOR ANALYSIS

>PROBLEM NITEM=30,RESPONSE=3;
>COMMENT;
   Análise de STEPWISE FULL INFORMATION FACTOR
 ANALYSIS,
   VARIMAX ROTATION AND PROMAX ROTATION.
   Os dados estão dispostos da seguinte maneira:
```

```
   Coluna 1 a 4: Sujeito
   Coluna 5: Sexo, onde 1 = masculino e 2 = feminino
   Coluna 6: Escolaridade (3 e 4: II Grau; 5 e 6: superior)
    Colunas 7 a 36: Respostas originais aos itens
   Colunas 37 a 66: Respostas corrigidas
>NAMES    VI1,VI2,    VI3,    VI4,    VI5,    VI7,
 VI8,VI9,VI10,VI11,VI12,
 VI13,VI14,VI15,VI16,VI17,VI18,VI19,VI20,VI21,VI22,
VI23,VI24,VI25,VI26,VI27,VI28,VI29,VI30;
>RESPONSE ' ','0','1';
>KEY 111111111111111111111111111111;
>RELIABILITY KR20;
>TETRACHORIC RECODE,NDEC=3,LIST;
>FACTOR
 NFAC=2,NROOT=3,NIT=(5,0.01),ROTATE=PROMAX,RESIDU
 AL,SMOOTH;
>FULL ITER=(20,3,0.005),OMIT=RECODE,STEPWISE;
>SAVE SMOOTH,ROTATED,TRIAL;
>INPUT NIDW=36,SCORES,WEIGHT=PATTERN;
(36X,I30)
>STOP  ;
```

Índice de assuntos