TEORIA CLASSICA DOS TESTES

TÓPICOS

1. Referências iniciais
2. Análise estatística dos itens de um teste
3. Análise Estatística de itens num teste usando SPSS
4. Alcances finais
5. Tarefa
6. Referencias

1. Referências iniciais

1. Gulliksen, H. (1950). Theory of Mental Tests. New York: John Wiley and Sons.

2. Lord, F.M., Norvick, M.R. (1968). Statistical Theories of Mental Test Score. Reading: Addison-Wesley.

3. Vianna, H.M. (1987). Testes em Educação. São Paulo: Ibrasa.

2. Análise estatística dos itens de um teste

- O objetivo geral em construção de testes é a obtenção de um teste de tamanho mínimo que produz escores confiáveis e válidos para o uso que se desejar dar a ele. Isto normalmente é alcançado testando um grande número de itens e selecionando aqueles que mais contribuem para a validade e confiabilidade do teste.

- Esses itens são identificados através do processo chamado análise de itens. Os parâmetros dos itens geralmente examinados são: função de variância e correlação com o critério.

- A análise estatística dos itens tem por finalidade estudar o comportamento psicométrico tanto de cada um de eles como de todo o conjunto. Usando vários métodos estatísticos e fazendo uso da interpretação da informação

nos permite garantir a validade e confiabilidade do instrumento que é construído.

2.1 Tamanho da amostra

- Com respeito ao tamanho da amostra Lazarte (1995) diz que "não existe uma regra absoluta sobre o tamanho da amostra. Certamente, uma análise de itens para testes nacionais envolve uma amostra grande e obtida cuidadosamente.

- Para teses e outros trabalhos feitos por alunos recomenda-se amostras nos 200". Uma regra empírica recomendada por Nunnally (1987) é usar entre 5 e 10 indivíduos para cada item no teste a ser analisado.

- Devemos levar em conta que uma amostra adicional será necessária para estudar a validação cruzada da que falaremos logo.

2.2 Etapas da análise estatística dos itens

- Segundo Ezcurra (1995), os passos a considerar numa análise estatística dos itens são:

– Selecionar uma amostra representativa de indivíduos em que é aplicado o teste piloto, que deve ser pelo menos igual ou maior do que os 200 casos, se o teste é para uma pesquisa e 1000, se o teste é para uso comercial.
– Qualificar os testes de acordo com a grelha de correção.
– Preparar o banco de dados, utilizando o seguinte modelo:

Tabela 1. Banco de dados típico para análise estatística dos itens

Pessoas	Itens						
	I1	I2	. . .	Ij	. . .	Ik	
P1	X_{11}	X_{12}		X_{1j}		X_{1k}	
P2	X_{21}	X_{22}		X_{2j}		X_{2k}	
.	.	.	.	.		.	
.	.	.	.	.		.	
.	.	.	.	.		.	
Pi	X_{i1}	X_{i2}		X_{ij}		X_{ik}	
.	.	.	.	.		.	X_{n1}
.Pn	X_{n2}		X_{nj}		X_{nk}		

Onde:
X_{ij} representa o valor ou escore obtido pelo indivíduo i no item j, que pode ser dicotômico ou policotômico.

2.3 Processo de Análise de Itens

- Lazarte (op. cit) considera o seguinte processo numa análise de itens:
  - Decidir quais propriedades do escore total são importantes (ou seja, maximizar a variabilidade, maximizar a predição de critérios externos, etc).
  - Identificar os parâmetros dos itens mais relevantes para estas propriedades do escore total.
  - Aplicar os itens para uma amostra de examinados que seja semelhante à população para a qual o teste está sendo construído.
  - Estimar os parâmetros dos itens especificados no passo 2.

- Estabelecer um plano para selecionar os itens ou identificar e revisar aqueles que estão com defeito.
- Selecione um grupo final de itens.
- Avaliar se o teste satisfaz o objetivo no passo 1, utilizando um estudo de validação cruzada.

2.4 Processamento dos dados psicométricos dos itens

- O processamento dos dados para obter as propriedades psicométricas dos itens, de acordo Ezcurra (op. cit.) envolve realizar os seguintes tipos de análise de forma obrigatória:
  - A distribuição de frequência dos escores totais e de cada subteste (se o teste tem subtestes).
  - Representar graficamente (polígonos de frequência ou histogramas) as distribuições de frequência dos escores totais e de cada subteste.

- Calcular a média, variância, desvio padrão, assimetria e curtose da distribuição dos escores totais e parciais de cada subteste.
- Obter a dificuldade do item (as proporções da resposta correta para cada item), e corrigir para evitar o efeito do acaso, assim como a proporção da escolha de cada um dos distratores (outras alternativas de resposta propostas) incluídos.
- Calcular a variância e desvio padrão de cada item, assim como a média e o desvio padrão do escore total e dos escores parciais dos indivíduos que escolheram a resposta correta.
- Calcular a dificuldade de cada item.
- Calcular o poder discriminativo de cada item.
- Calcular o coeficiente de validade de cada item.

Opcional:

- Calcular a matriz de correlação entre os sub-testes, e entre o escore total e cada sub-teste.
- Calcular a análise de regressão múltipla dos sub-testes, e sob o escore total de modo que a partir da estimação dos coeficientes de regressão parcial possa-se fazer o peso para cada sub-teste.
- Calcular a análise fatorial da matriz de intercorrelação dos itens para estabelecer a existência de fatores comuns.

2.5 Tipo de análise

Os principais tipos de análise estatística utilizados hoje de preferência nos testes de desempenho, atuação ou aptidão (Nuria Cortada Kohan, 1968; Magnusson, 1990; Kline 1986, Nunnally, 1987), são:

2.5.1 Dificuldade do Item, média e variância

Itens dicotômicos

São os mais comuns nos testes de aptidão. O item médio corresponde à proporção de examinados que responderam o item "corretamente". Para o item i essa proporção, p_i, é chamada de dificuldade do item ou índice de dificuldade. Também pode ser apresentada como o percentual de pessoas que responderam corretamente o item através de:

$$Dif = \frac{Número\ de\ indivíduos\ que\ responderam\ corretamente\ o\ item}{Número\ de\ indivíduos\ availados} \times 100$$

Correção da dificuldade para o acaso:

- Quando o teste de desempenho é do tipo múltipla escolha, é necessário corrigir seu valor, devido à provável existência da adivinhação nas respostas dos indivíduos. Este procedimento consiste na correção de aleatoriedade, e é calculada a partir da seguinte fórmula (Guilford, 1954; Magnusson, 1990):

$$P = \frac{R - \frac{W}{0-1}}{N}$$

Onde:

P = Dificuldade corrigida.

R = Número de indivíduos que marcaram corretamente o item.

W = Número de indivíduos que marcaram incorretamente o item.

O = Número de alternativas que têm o item
N = Número total de indivíduos avaliados.

- Assim, em uma amostra, a dificuldade do item pode ser afetada pelo formato do item. Se o formato é de múltipla escolha, como dizemos, o número de alternativas pode influenciar p_i. Por exemplo, se 50% realmente sabe a resposta correta, e a outra metade adivinha, teríamos as seguintes proporções:

Nº Alternativas	Proporção que adivinha	p_i
5	0.5/5 = 0.100	0.600
4	0.5/4 = 0.125	0.625
3	0.5/3 = 0.167	0.675
2	0.5/2 = 0.250	0.750

- Estas proporções podem ser ainda maiores se considerarmos que a resposta correta pode ser obtida se algumas alternativas obviamente erradas são eliminadas. Em muitos testes de aptidão usados nos EUA as dificuldades do item reportadas variam geralmente entre 0,6 e 0,8, em parte devido a esse fenômeno de adivinhar.

- Portanto, para itens de múltipla escolha é aconselhável obter, para além da média e da variância do item, a distribuição de frequências para as alternativas que foram escolhidas pelos avaliados. As alternativas que não são a resposta correta são chamadas distratores. Esta distribuição pode indicar se existem distratores que não atraem nenhuma resposta, ou que atraem a maioria das respostas sem ser a correta, etc.

- Por exemplo, na tabela adjacente o item 1 é difícil, porque um dos distratores atrai a maioria dos indivíduos. No item 2, dois distratores não

funcionam em absoluto. No item 3, temos o caso clássico de um item com distratores aceitáveis.

	Alternativas (%)				
Item	A	B	C	D	p_i
1	24	4	56	16*	0,16
2	92*	0	8	0	0,92
3	20	20	8	52*	0,52

- Depois de corrigir a dificuldade é possível ordenar os itens de mais fácil para o mais difícil, como é o caso em testes de dificuldade crescente, no caso de testes de poder é recomendado selecionar os itens com níveis de dificuldade entre 0,50 e 0,60.

- Para os itens dicotômicos, a variância da amostra do item deve ser descartada pois não fornece informação sobre as diferenças entre os avaliados.

- Um item oferece a maior quantidade de informação sobre as diferenças entre os avaliados, quando $p_i = 0.5$ (Dif = 50%), e portanto a variância é maximizada.

- Por isso, recomenda-se selecionar os itens em um intervalo de cerca de 0.5 (alguns autores sugerem entre 0,3 e 0,7).

- Se o teste é para selecionar indivíduos, os itens mais difíceis são recomendados.

Tabela 2. Classificação do nível de dificuldade dos itens dicotômicos *

CLASSIFICAÇÃO	ÍNDICE DE DIFICULDADE
MUITO FÁCIL	DE 0.75 A 0.99
FÁCIL	DE 0.55 A 0.74
INTERMEDIÁRIO	DE 0.45 A 0.54
DIFÍCIL	DE 0.25 A 0.44
MUITO DIFÍCIL	DE 0.05 A 0.24

* Tomado de Ezcurra (op. cit)

Itens Poliatômicos

Os mais comuns nas escalas de Atitudes. Neste caso é requerido obter independentemente a média e a variância dos itens. A media é equivalente de p_i nos itens dicotômicos, pero agora não tem interpretação de dificuldade. A variância dos itens nos ajuda a escolher aqueles itens no sentido que procuramos aqueles com a maior variância possível.

2.5.2. Discriminação do item

- Mede o grau em que o item é capaz de estabelecer diferenças entre os indivíduos com altos e baixos níveis de uma habilidade, aptidão ou conhecimento que está sendo avaliado.

- O objetivo de qualquer teste é fornecer informação sobre as diferenças individuais no construto medido pelo teste, ou num critério externo, que o

teste supostamente prediz. Portanto, estamos interessados em obter índices que mostram como efetivamente um item discrimina entre os avaliados que têm altos escores no critério e aqueles que têm baixos escores.

- Na ausência de um critério externo, o escore total do mesmo teste é utilizado. Assim, o objetivo é identificar itens que os indivíduos que tem altos escores respondem corretamente com uma alta probabilidade, enquanto que os indivíduos com baixos escores respondem incorretamente.
- Um item que é respondido igualmente de forma correta por indivíduos com escores altos e baixos, não discrimina bem entre esses dois grupos e não seria útil.

- Um item que é respondido corretamente pelos indivíduos de escore baixo, e incorretamente pelos de alto escore, é um item com a discriminação negativa e não é desejável.

Índice de Discriminação

- Este índice aplica-se só aos itens dicotômicos. Determina-se na distribuição dos escores do critério, um ou dois pontos de corte e classifica-se aos avaliados em grupos com escores abaixo e acima desses pontos de corte. Por exemplo, dividir em duas metades e classificar indivíduos na metade inferior e superior, dividir no terço superior e o terço inferior, etc.

- Por exemplo, no seguinte:

- Grupo superior, que representa o 27% dos casos com escores totais maiores.
- Grupo intermediário, que representa o 46% dos casos com escores intermediários.
- Grupo baixo, que representa o 27% dos casos com escores totais menores.
- Deles separam-se os grupos extremos

- Uma vez que os dois grupos foram identificados, o índice de discriminação, D_i, do item I é obtido como: $D_i = p_{i\ S} - p_{i\ I}$ onde p_{iS} é a proporção de indivíduos no grupo superior que respondeu o item corretamente, e $p_{i\ I}$ é a proporção de corretas do grupo inferior.

- De outra forma como regra geral, no grupo superior e no grupo inferior, são calculados separadamente para cada item a percentagem de

indivíduos que responderam corretamente, ambos dados são subtraídos e o resultado final é a discriminação que têm o item, sua fórmula é:

$$\text{Disc.} = \% \text{ correto do grupo Superior} - \% \text{ correto do grupo Inferior}$$

- Disc. pode variar entre -1 e 1. Os valores positivos indicam que o item discrimina em favor do grupo superior, os negativos indicam que o item é discriminador que favorece ao grupo inferior.

- Uma regra prática para avaliar D_i segundo Lazarte (Op. Cit.) é:

  D = 0.40 o item funciona bem

  0.30 = D = 0.39 pouca ou nenhuma revisão é requerida

  0.20 = D = 0.29 o item é marginal, precisa de revisão

D = 0.19 o item deve ser removido ou substituído

- Esta regra pode ser resumida na tabela seguinte:

Tabela 3. Classificação da discriminação dos itens dicotômicos *

CLASSIFICAÇÃO	DISCRIMINAÇÃO
MUITO BOA DISCRIMINAÇÃO	DE 0.40 A 0.99
DISCRIMINAÇÃO ACEITÁVEL	DE 0.30 A 0.39
DISCRIMINAÇÃO INTERMEDIÁRIA	DE 0.20 A 0.29
DISCRIMINAÇÃO INACEITÁVEL	DE 0.05 A 0.19

* Tomado de Ezcurra (op. cit.)

2.5.3 Validade do item

Mede o grau no qual um item mede validamente aquela capacidade que deseja-se medir.

Índices de correlação de validação do item

- Todos esses índices correlacionam o escore no item com o escore obtido no critério externo, ou, na ausência de critérios externos, o escore total obtido no mesmo teste.

- Em geral, todos esses índices são chamados *correlações item-total*. Quando o item é policotômico (como um item Likert), a correlação entre o item e o total é a correlação de Pearson entre outros casos receberam novos nomes, como veremos logo.

- Ao usar o escore total do mesmo teste como critério, as correlações são modificados para eliminar a contribuição ao escore total do item estudado. Este tipo de correlação é chamado *correlação item-total com o item removido.*

- Geralmente os coeficientes de correlação item-teste são utilizados para quantificar, os mais usados são:

Correlação r de Pearson

É usada em situações em que as duas variáveis correlacionadas são contínuas. Utiliza-se a seguinte fórmula:

$$\rho_{i\,X} = \frac{\sigma_{i\,X}}{\sigma_i \sigma_X},$$

Para corrigir o resultado utiliza-se a seguinte fórmula:

$$\rho_{\,i\,(X-i\,)} = \frac{\rho_{iX}\sigma_X - \sigma_i}{\sqrt{\sigma_X^2 + \sigma_i^2 - 2\rho_{i\,X}\sigma_X\sigma_i}}$$

Onde:

$\rho_{\,i\,(X-1)}$ = Correlação corrigida item-teste.

$\rho_{\,i\,X}$ = Correlação item-teste.

σ_X = Desvio padrão dos escores totais dos indivíduos avaliados.

σ_i = Desvio padrão dos escores do item.

σ_{iX} = Covariância entre o item e o escore total.

Quanto mais próximo o coeficiente é de 1 é melhor, e aceita-se como critério empírico para aceitar o item que o resultado obtido deve ser, pelo menos, superior ou igual a 0.20.

Correlação Bisserial

Usada em situações em que uma variável que se correlaciona é contínua e a outra é dicotômica. É a correlação produto-momento de Pearson entre uma variável dicotômica (0 ou 1) e uma variável contínua. É o caso típico de itens dicotômicos. A fórmula para calcular essa correlação é dada por:

$$\rho_{\rho bis} = \frac{\mu_{i+} - \mu_X}{\sigma_X}\sqrt{p_i\ q_i}$$

Onde:

μ_{i+} = Média no critério (a média dos escores totais) dos indivíduos que respondem corretamente o item i.

μ_X = Média ou média dos escores totais de todos os indivíduos no teste.

σ_X = Desvio padrão dos escores totais dos indivíduos avaliados.

p_i = Proporção de indivíduos que respondem corretamente o item i. (Dificuldade do item i)

Quanto mais próximo o resultado é do valor 1, o coeficiente será melhor, e aceita-se como critério empírico que este deve ser, pelo menos, superior ou igual a 0.20 para ter em conta o item.

Na maioria dos casos para calcular o escore total e analisar um item, o resultado do mesmo está incluído no escore total, se o número de itens é grande (25 ou mais), isso não é um problema.

Se não for o caso, é necessário corrigir esta situação pois introduze ao resultado final um aumento do mesmo por efeito da autocorrelação, em geral pode-se corrigir a correlação removendo o item do total utilizando a seguinte fórmula:

$$\rho_{\text{pbis c}} = \frac{\rho_{\text{pbis}}\sigma_X - \sigma_i}{\sqrt{\sigma_X^2 + \sigma_i^2 - 2\rho_{\text{pbis}}\sigma_X\sigma_i}}$$

Aqui p_{pbis} é a correlação bisserial original entre o item e o escore total do critério, s_x é o desvio padrão do item, e $p_{ibis\,c}$ é a correlação bisserial corrigida quando o item i é removido do escore total.

Note-se que esta equação pode ser aplicada a qualquer tipo de correlação original, e não apenas à ponto-bisserial.

Coeficiente Phi

Quando os itens dicotômicos devem correlacionar-se com os critérios dicotômicos (ingressante vs. não ingressante, sucesso vs. falha, etc.), a extensão da correlação produto-momento de Pearson é chamada coeficiente Phi.

Como a covariância entre os itens i e o critério dicotômico X, e suas respectivas variâncias são uma função da proporção de indivíduos que passam o item, p_i, e a proporção de indivíduos que passam o critério, p_x, é possível mostrar que o coeficiente Phi pode ser expressado como:

$$p_{iX}\,(\text{phi}) = \frac{p_{iX} - p_i p_X}{\sqrt{p_i(1-p_i)p_X(1-p_X)}}$$

Onde P_{iX} é a proporção de indivíduos que passam o item, e também passam o critério; P_i é a proporção de indivíduos que passam o item i, e P_x é a proporção de indivíduos que passam o critério.

2.6 Índices de Confiabilidade e Validez do item

- Os índices confiabilidade e validez do item são funções conjuntas da variância do item e de sua correlação com o critério.

- Se o critério usado é o escore total na mesma prova (critério interno) o índice se denomina *índice de confiabilidade* do item e se define como

$$\sigma_i \rho_{iX}$$

em que σ_i é o desvio padrão do item e ρ_{iX} é a correlação item-total.

Quando um critério é usado, o índice se denomina *índice de validez* do item e se define de modo similar como

$$\sigma_i \rho_{iY},$$

em que ρ_{iY} é a correlação entre o item e um critério externo.

- Estes índices são úteis pois sua combinação aditiva gera a variância do escore total, e o coeficiente de validez entre o teste e um critério externo pode ser expresso como a razão da suma dos índices de confiabilidade e validez, isto é:

$$\sigma_X^2 = \left(\sum_{i=1}^{k} \sigma_i \rho_{iX} \right)^2 , \qquad \rho_{XY} = \frac{\sum_{i=1}^{k} \sigma_i \rho_{iY}}{\sum_{i=1}^{k} \sigma_i \rho_{iX}}$$

- O índice de confiabilidade do item pode ser utilizado para estimar o valor do coeficiente alfa de Cronbach quando um novo item é retirado do teste. A expressão a usar é:

$$\alpha = \frac{k}{k-1} \left[1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\left(\sum_{i=1}^{k} \sigma_i \rho_{iX} \right)^2} \right]$$

em que k representa o número de itens selecionados para entrar no teste até esse momento.

2.7 Crosvalidation ou Validação cruzada

- Quando os itens são selecionados sobre a base de critérios estatísticos usando as respostas de uma amostra dada, o teste assim construído deveria ser muito efetivo para essa amostra em particular, mas não necessariamente em uma outra amostra.

- Num estudo de validação cruzada o criador do teste usa itens que tem escolhido considerando uma análise de itens, estes itens são aplicados a uma segunda amostra, independente da primeira, e a confiabilidade e validez dos escores são avaliados de novo.

- Para obter informação relevante numa aplicação só do item, a amostra original - a qual se aplica todos os itens - é dividida em dois grupos aleatoriamente. Num grupo é feito a análise dos itens. Logo, no outro grupo, analisasse o escore total do teste baseado nos itens selecionados na análise dos itens do primeiro grupo.

- Quando as análises dos itens no primeiro grupo se usam para escolher itens no segundo grupo, e ainda os resultados dos itens do segundo grupo se usam para selecionar os itens do primeiro grupo falamos de validação cruzada dupla.

2.8 Critérios para a escolha dos itens

- Finalmente quando todos as análises estatísticas dos itens são completadas precisasse de uma revisão crítica dos mesmos.

- Esta revisão deve ser feita considerando:

– Analisar a dificuldade de cada um dos itens de modo que posa se formar grupos de dificuldade e fazer uma ordem entre eles.
– Analisar a discriminação dos itens e retirar aqueles que tenham valores muito baixo, inferiores ao critério empírico recomendado.
– Analisar a validez dos itens, removendo aquele que não satisfazem o critério mínimo considerado.
– Analisar para cada item de forma conjunta a dificuldade, discriminação e os outros critérios, e então escolher aqueles que satisfazem os três critérios ou boa parte deles ao mesmo tempo.

- Geralmente logo de fazer as análises de itens quando construísse um teste pela primeira vez, são removidos uma grande quantidade de itens, porem precisasse de que no piloto sejam aplicadas uma grande quantidade dos mesmos.

- Mas se acontece que o número de itens fica pequeno, então precisasse fazer itens adicionais e aplicar a uma nova amostra e volver a fazer os análises apresentados.

- A ideia final e obter um teste com o qual obter a versão definitiva da validez e confiabilidade do teste e ainda estabelecer tabelas de interpretação.

3. Análise Estatística de itens num teste usando SPSS

- A análise estatística dos itens de um teste, é uma etapa da metodologia de construção do teste, consistente na utilização de critérios e técnicas para a eliminação de itens. A análise estatística dos itens sempre precede à estimação dos parâmetros de um teste pois estes são obtidos com uma versão chamada Teste Final.

- A metodologia estatística baseada nas etapas da construção de testes apresenta-se desde o ponto de vista sequencial, dessa forma a análise estatística dos itens permite uma análise psicométrica preliminar pois elimina itens que contribuem pouco à qualidade de ajustamento dos parâmetros de teste: tanto na diferenciabilidade, a validade de construto como na confiabilidade.

- O objetivo geral em construção de testes é obter um teste de poucos itens que produz pontuações válidas e confiáveis para o uso que deseja-se dar. (Lazarte, 1995)

- Este processo pode ser obtido adequadamente a partir do módulo de confiabilidade do SPSS mas ele está disponível em diferentes pacotes, especialmente com estatísticas de item e do teste, assim como do teste se o item foi excluído, covariância e correlações inter-item, e resumos das médias, variâncias, covariâncias e correlações. Especialmente o cálculo do alfa de Cronbach.

3.1 Tipos de análises estatísticas

3.1.1 Média e Variância

- Aqui é necessário para obter independentemente a média e variância dos itens. A média é utilizada para estabelecer a homogeneidade dos itens. A variância ajuda a selecionar itens, no sentido de encontrar aqueles itens com a maior variância possível.

3.1.2 Índices Correlacionais de validação do item

- Todos esses índices correlacionam o escore no item com o escore obtido no critério externo, ou, na ausência de critério externo, o escore total obtido no mesmo teste.

- Em geral, todos esses índices são chamados correlações item-total. Quando o item é policotômico (como um item de Likert), a correlação entre o item e o total é a correlação de Pearson.

- Ao usar o escore total do mesmo teste como critério, as correlações (correlação espúria) são modificadas para eliminar a contribuição ao escore total do item estudado.

- Essa correlação é chamada correlação ítem-total com o item removido ou eliminado. A correlação corresponde ao r de Pearson cuja fórmula é conhecida:

$$\rho_{i\,X} = \frac{\sigma_{i\,X}}{\sigma_i \sigma_X},$$

e para corrigir os resultados utiliza-se a seguinte fórmula:

$$\rho_{\,i\,(X-i\,)} = \frac{\rho_{iX}\sigma_X - \sigma_i}{\sqrt{\sigma_X^2 + \sigma_i^2 - 2\rho_{i\,X}\sigma_X\sigma_i}}$$

Onde:

$\rho_{i(X-1)}$ = correlação item-teste corrigida.

ρ_{iX} = correlação item-teste.

σ_X = desvio padrão dos escores totais dos indivíduos examinados.

σ_i = Desvio padrão dos escores do item.

σ_{iX} = Covariância entre o item e o escore total.

- É melhor quanto mais próximo o coeficiente é de 1, e aceita-se como critério empírico para aceitar o item que o resultado deve ser, pelo menos, superior ou igual a 0,20.

3.1.3 Índices de Confiabilidade do item

- Os índices de confiabilidade de itens são funções conjuntas da variância do item e de sua correlação com o critério. Se o critério utilizado é o escore total no mesmo teste (critério interno) o índice é chamado índice de confiabilidade do item, e define-se como $\sigma_i\rho_{iX}$, onde σ_i é o desvio padrão do item e $\rho_{i\,X}$ representa a correlação item-total.

- Esses índices são úteis porque sua combinação aditiva gera a variância do escore total,

$$\sigma_X^2 = \left(\sum_{i=1}^{k}\sigma_i\rho_{i\,X}\right)^2 \quad ,$$

- O índice de confiabilidade do item também pode ser usado para re-estimar o valor do coeficiente alfa de Cronbach quando um novo item é adicionado ao teste, ou subtraído do teste. A fórmula utilizada é a seguinte:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i\rho_{iX}\right)^2}\right]$$

onde k é o número de itens selecionados para entrar no teste até esse momento.

- Geralmente, depois da realização dessas análises, eliminam-se muitos itens, por isso é necessário que o teste inicial tenha muitos deles.

- Se for o caso, que o número de itens a ser retida é muito pequena, devem criar-se novos itens adicionais e aplicá-los a uma amostra e realizar a análise acima descrita, de modo que possa-se construir a versão final que será utilizada para o estudo da validade, confiabilidade e Diferenciabilidade de um teste de uma forma mais elaborada.

3.2 Uso do Módulo de Confiabilidade do SPSS para a análise estatística de itens do teste

- A análise estatística dos itens do teste pode ser obtido usando o método alfa do módulo de confiabilidade do software estatístico SPSS. Um importante relatório sobre este método são as "Estatísticas item-total".

- Os relatórios dessas estatísticas item-total são média e variância do item, a média e a variância do teste se o item foi eliminado, a correlação item-total corrigida (excluindo o item no teste), o alfa de Cronbach para o teste se o item foi eliminado.

- A média e o desvio padrão do item são estatísticas descritivas do item, a média e a variância do teste se o item foi eliminado são indicadores da diferenciabilidade do teste, a correlação item-total corrigida é um

indicador da validade do item e o alfa de Cronbach se o item foi eliminado é um indicador de validade.

- Na Tabela mostra-se um relatório de Estatísticas Item-Total para um teste de Escalas Likert em atitudes. Escala é de 1 a 5. Os altos valores expressam atitudes negativas.

Tabela 4. Estatísticas de Validade e Confiabilidade de Itens da EAHM

Item	Item	Media del ítem si el ítem se elimina	Ds. Es. del ítem corregida	Promedio de la prueba si el ítem se elimina	Varianza de la prueba si el ítem se elimina	Correlación ítem-prueba corregida	alpha si el ítem se elimina
	I01	3.0781	1.0561	110.8359	312.0436	.4585	.8925
	I02	3.1016	1.1296	110.8125	312.5922	.4106	.8933
	I03	3.7813	1.2738	110.1328	306.5470	.4954	.8917
	I04	2.5117	1.1614	111.4023	319.3473	.2306	.8964
	I05	3.2305	1.2546	110.6836	304.2720	.5579	.8905
	I06	3.9805	1.1042	109.9336	307.9995	.5431	.8910
	I07	3.0586	1.0666	110.8555	308.9084	.5393	.8912
	I08	2.9453	1.0977	110.9688	307.0735	.5716	.8905
	I09	3.9375	1.1044	109.9766	311.7014	.4448	.8927
	I10	3.4570	1.0982	110.4570	308.8923	.5225	.8914
	I11	3.2695	1.0105	110.6445	308.3398	.5891	.8905
	I12	3.1836	1.0411	110.7305	311.3114	.4865	.8921
	I13	2.7578	1.4072	111.1563	315.2225	.2614	.8967
	I14	3.1328	1.1263	110.7813	308.6186	.5150	.8915
	I15	2.9375	1.1931	110.9766	308.8465	.4768	.8921
	I16	3.7891	1.1213	110.1250	310.9412	.4569	.8925
	I17	3.7656	1.0174	110.1484	310.9112	.5106	.8917
	I18	3.8242	1.0349	110.0898	310.2625	.5193	.8916
	I19	2.8984	1.0652	111.0156	310.2586	.5029	.8918
	I20	3.1680	1.2861	110.7461	310.2765	.4047	.8935
	I21	3.7188	1.1057	110.1953	309.8127	.4941	.8919
	I22	4.2969	1.0393	109.6172	318.0176	.3013	.8950
	I23	2.1953	1.1030	111.7188	336.3363	-.1801	.9028
	I24	2.8477	1.1394	111.0664	325.3642	.0877	.8988
	I25	2.9063	1.1128	111.0078	309.0823	.5098	.8916

I26	3.1563	1.1778	110.7578	312.7960	.3861	.8937
I27	3.6445	1.0602	110.2695	315.6486	.3582	.8941
I28	3.1133	1.2138	110.8008	311.7209	.3983	.8935
I29	3.4453	1.0122	110.4688	307.6382	.6085	.8902
I30	3.4727	1.2010	110.4414	316.7103	.2834	.8956
I31	4.2734	.9918	109.6406	314.8350	.4105	.8933
I32	3.9727	1.2124	109.9414	306.3534	.5288	.8911
I33	3.8750	1.0738	110.0391	314.1161	.3941	.8935
I34	3.1875	1.0646	110.7266	308.8504	.5421	.8911

Puntaje de la Prueba
Media: 113.9141 Desv. Est.: 18.1733 Casos: 256
Variancia: 330.2688
Coeficiente de Confiabilidad 34 itemes Alpha = .8958

- Observa-se que os itens 6,22,31,32 e 33 têm média relativamente elevada em relação aos outros itens. Isto significa que a amostra de 256 testados tem uma maior tendência de avaliar negativamente as expressões desses itens. De outra forma, o item 23 expressaria uma maior tendência para a avaliação da expressão deste item como de atitude positiva.

- Com respeito à variabilidade pode-se dizer que os itens 13,20,30 e 32 têm maior variabilidade e o item 31 tem menor variabilidade.

- Com respeito ao escore do teste, este corresponde a uma média de cerca de 114 pontos e um desvio padrão de cerca de 18 pontos. Este escore pode variar de 34 pontos (escolhe 1 nos 34 itens) a 170 pontos (escolhe 5 nos 34 itens), como fora dito anteriormente.

- De acordo com a tabela, o alfa é alto e mostra que o teste é adequado. Az correlações item-total corrigido são de moderadas a elevadas. Apenas cinco itens (4, 13, 23, 24 e 30) não excedem o valor de 0,30, e podem ser eliminados.

- Os itens 23 e 24 devem ser removidos, eles mostram correlações item-total negativas e nulas, respectivamente, que podem ser interpretadas em termos de validade, pois eles não medem o construto avaliado. A remoção também leva a um ganho substancial de variabilidade e de alfa de Cronbach.

4. Alcances finais

- As estatísticas mais apropriadas apresentadas pelo módulo SPSS são a média do teste se o item foi eliminado, a variância do teste se o item foi eliminado, a correlação item-teste corrigida e o alfa se o item é eliminado. Porém, também é considerado conhecimento do pesquisador para decidir quais itens serão eliminados. Por isso, afirma-se que a análise estatística dos itens, consiste em técnicas, mais ou menos adequados.

- Um aspecto não desenvolvido neste estudo é a regressão do item com o teste. Esta técnica é utilizada em alguns relatórios de psicologia, não demonstrou sua relevância e sua justificação neste estudo, mas merece um estudo mais detalhado para o caso de itens dicotômicos.

- O módulo de confiabilidade do SPSS que está disponível em outros pacotes incluindo o R é mais adequado para a análise estatística dos itens,

especialmente com o cálculo da média, variância e alfa de Cronbach se o item é eliminado, e a correlação item-total corrigida.

- Na prática as diferentes medidas estatísticas do análise de itens são usados para decidir se o item permanece como um elemento do teste ou não. Um exemplo sobre o uso destes métodos está na seguinte referência.

Bazán, J. L., Millones, O. (2002b). *Evaluación psicométrica de las pruebas CRECER 98*. En: Rodríguez, J., Vargas, S. (eds). Análisis de los Resultados y Metodología de las Pruebas Crecer 1998. Documento de trabajo 13. Lima: MECEP-Ministerio de Educación. Pp: 171-195.

https://jorgeluisbazan.weebly.com/uploads/1/2/5/6/125695412/13h.pdf

5. Tarefa

1. Fazer uma análise de itens considerando as diferentes propostas usando os dados de Prova de conhecimentos em Matemática: mathbfinal.sav

2. Fazer uma análise de itens considerando as diferentes propostas usando os dados de atitudes frente a Estatística: baseunionfinal.sav

3. Use diferentes pacotes do R como CTT e psych.

6. Referencias

1. EZCURRA, L. (1995) Análisis Estadísitico de Items. Separata del curso Seminario de Construcción de Pruebas I. UNMSM. Facultad de Psicología. 3 p.

2. GUILFORD, J. P. (1954) *Psychometrics Methods*, New York Mc Graw Hill.

3. KLINE P. (1986) *A Handbook of Test Construction: Introduction to Psychometric Desig*, New York, Methuen And. Co., Ltd.

4. LAZARTE, A. (1995) Análisis de Items. Separata del curso PSB234. PUCP. Facultad de Psicología 3p.

5. MAGNUSSON, D. (1990) *Teoría de los Test.* Edit. Trillas México

6. NUNNALLY, J. (1987) *Teoría Psicométrica*, México. Ed. Trillas.

7. NURIA CORTADA DE KOHAN (1968) *Estadística Aplicada.* Bs. Aires. Argentina.