Minicurso

MC-12 - AVALIAÇÃO EDUCACIONAL: ENTENDENDO A TEORIA DA RESPOSTA AO ITEM (ABE)

Ministrantes: Jorge Luis Bazán (USP) e Mariana Curi (USP)
jlbazan@icmc.usp.br, mcuri@icmc.usp.br

Público alvo: Professores do ensino básico

Sala: *AT 04 - Sala 82*
De 14/7/2015 à 17/7/2015 - das 08h00 às 10h00

AULA 2. ELABORAÇÃO DE QUESTÕES OU ITENS E SUA ANÁLISE USANDO METODOLOGIA TRADICIONAL (Oficina)

Conteúdo. *Principais medidas e critérios para a interpretação dos resultados.* Matrizes de referência. Tipos de itens. Recomendações para elaboração de itens. Exemplos. Oficina de redação de itens. Análise clássica de itens (qualitativa e quantitativa) baseada na TCT. Respostas correta, não resposta e valores perdidos. Identificação de viés. Principais medidas e critérios para a interpretação dos resultados.

Ministrante: Jorge Luís Bazán

TÓPICOS

1. MARCO METODOLÓGICO DA AVALIACAO
2. MATRIZES DE REFERENCIA OU MAPA DO CONSTRUTO
3. ELABORAÇÃO DE ITENS OU PLANEJAMENTO DA MEDIDA
4. ANALISE CLASSICA DE ITENS OU MODELO DE MEDIÇÃO CLASSICO
5. CRITERIOS PARA INTERPRETACAO DE RESULTADOS OU DO ESPAÇO DE RESULTADOS
6. ANALISIS DE ITENS USANDO SOFTWARE

1. MARCO METODOLÓGICO DA AVALIAÇÂO

Nos podemos considerar como marco metodológico uma versão adaptada da proposta de Duckor, Draney e Wilson (2009) também discutido em Wilson (2005).

Estes autores apresentam uma proposta para a construção de medidas com base em quatro etapas e princípios do sistema de avaliação, os quais são apresentados na Figura 1.

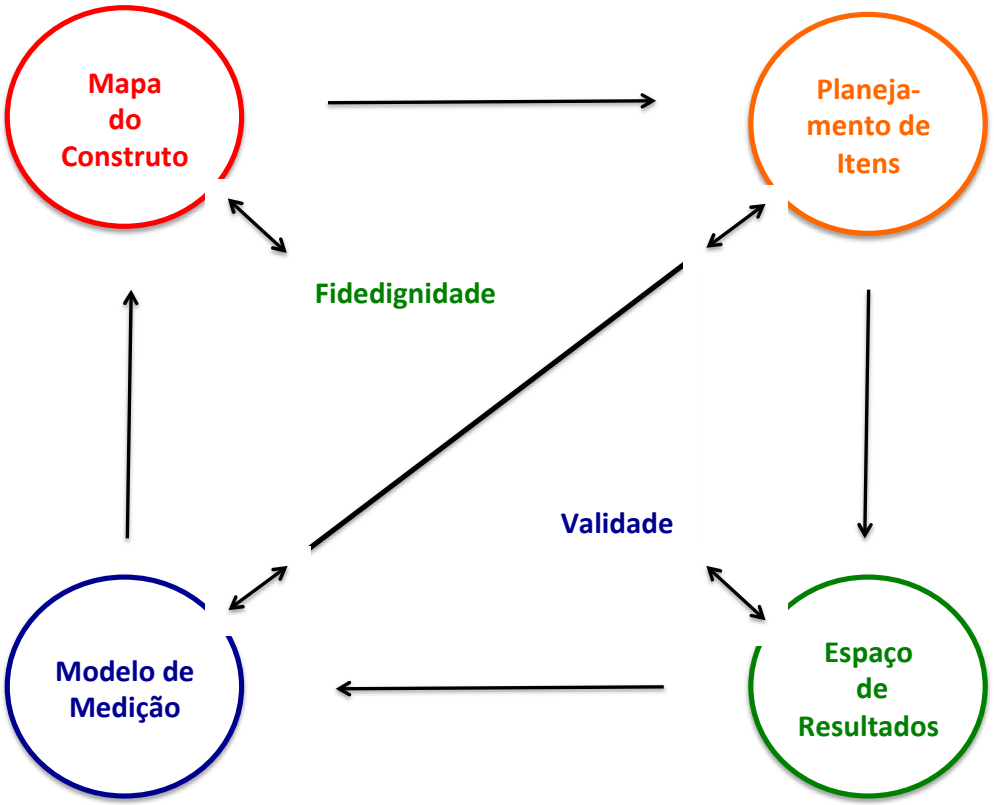


Figura 1. Relações entre os quatro blocos para a construção de medidas (tomado de Duckor, Draney e Wilson, 2009).

O processo da construção de uma medida inicia-se na 1) definição do mapa do construto, segue com o 2) planejamento de itens, a 3) definição do espaço de resultados, o qual define, e finalmente 4) o modelo de medição a ser considerado. Neste processo as Etapas 1-3 envolvem a fidedignidade[1] da medida enquanto que as Etapas 2-4 envolvem a validade[2] da mesma.

[1] A fidedignidade é uma característica da medida que faz referencia ao grau de consistência ou reprodutibilidade das medidas quando os procedimentos das avaliações são replicados sob as mesmas condições.

[2] A validade da medida faz referencia ao grau pelo qual a evidência e a teoria suportam as interpretações a partir dos valores das medidas.

Na pratica esse processo não necessariamente é explicito, isto é, os elaboradores ou construtores de medidas não necessariamente seguem esse processo no nível de detalhe discutido na proposta dos autores. Entretanto, quando se deve analisar uma medida é requerido avaliar cada uma dessas etapas. .

A analise de toda medida enfatiza tópicos que envolvem diferentes objetivos como: revisão, descrição, crítica ou proposta. Neste documento, propomos a classificação destes diferentes objetivos, dependendo do foco em que eles se centram. Por exemplo alguns trabalhos enfatizam sua revisão, descrição, crítica ou proposta na definição do mapa de construto, mas outros podem ser melhor classificados como centrando seus objetivos no modelo de medição.

Note em nossa proposta, trocamos a ordem das Etapas 3 e 4. Isto é, uma vez que as medidas já estão definidas, primeiro revisamos o modelo de medição e em seguida o espaço de resultados adotado. Neste caso, a sequencia fica semelhante a um planejamento estatístico usual.

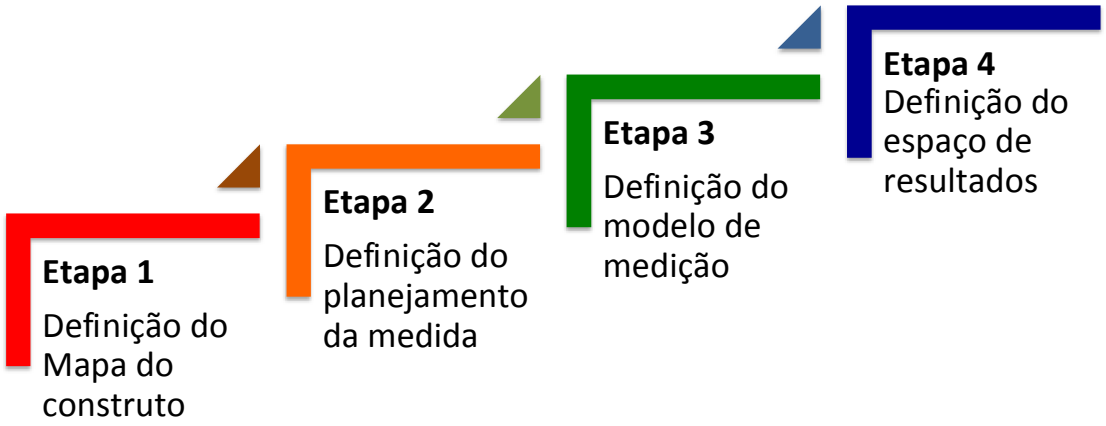


Figura 2. Etapas para a avaliação de construção de medidas (adaptado de Duckor, Daney e Wilson, 2009).

Quadro 1. Etapas de avaliação do processo de construção de medidas.

Etapa	**Nome da Etapa**	**Definição**	**Pergunta**	**Planejamento estatístico**
I	A definição do Mapa do construto	definição de aquilo que está sendo medido.	que vai ser medido?	Definição da pesquisa ou interação entre o pesquisador e o analista de dados
II	A definição do Planejamento da medida	definição do formato de avaliação ou instrumento e as unidades de observação o fontes de informação (alunos, diretores, etc.) processo, amostragem ou instrumentos, bases de dados	Como vai ser medido?	Definição dos instrumentos, amostragem, processo de captura de dados, elaboração de base de dados

Etapa	Nome da Etapa	Definição	Pergunta	Planejamento estatístico
III	Modelo de medição	definição do modelo de medição (modelo estatístico) que é aplicado em II		Definição do modelo estatístico ou técnica de análise de dados a serem adotadas
IV	definição da apresentaçã o do espaço de resultados	definição da forma de apresentação dos resultados finais e sua interpretação e uso que é aplicado ao processo em III.		Definição do modelo de reporte de resultados

2. MATRICES DE REFERENCIA O MAPA DO CONSTRUTO

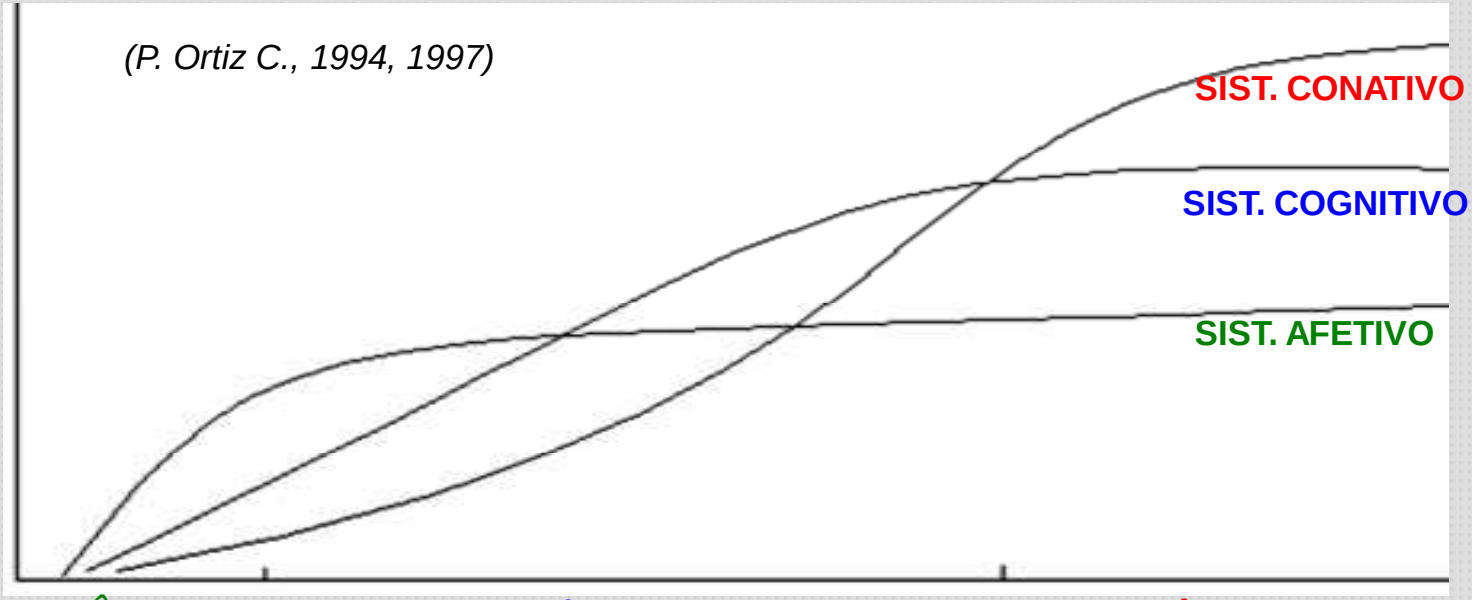
Mapa do construto

Um instrumento é sempre secundário. Há sempre uma finalidade para a qual é necessário um instrumento e do contexto em que este será utilizado (ou seja, envolvendo algum tipo de decisão).

Trata-se de uma idéia ou conceito que é o objeto teórico do nosso interesse em o avaliado conhecido comunmente como *construto* .

O construto pode ser parte de um modelo teórico de uma cognição pessoal - como a sua compreensão de determinado conjunto de conceitos ou atitude em relação a alguma coisa - ou pode ser alguma outra variável psicológica, ou o desempenho de determinado domínio educacional , etc. Ou pode estar relacionada com um grupo , em vez de um indivíduo ou podem ser um objecto inanimado complexo .

O DESENVOLVIMENTO DA CONSCIÊNCIA E A PERONALIDADE REFLETE A HISTORIA DA



INFÂNCIA 1:

OS SENTIMENTOS REFLETEM A ESTRUTURA TRADICIONAL

INFÂNCIA 2:

OS CONHECIMENTOS REFLETEM A ESTRUTURA CULTURAL

ADOLESCÊNCIA:

AS MOTIVAÇÕES REFLETEM A ESTRUTURA ECONÓMICA

Há uma infinidade de teorias - o importante aqui é ter uma estrutura para proporcionar motivação e estrutura para o construto a serem medido.

A idéia de construir um mapa é um conceito mais preciso do que falar de construto.

Supõe-se que o construto a ser medido tem uma forma particularmente simples, estendida de um extremo a outro, de alto ao um baixo valor, a partir de um pequeno a grande valor, de positiva para negativ, ou de forte para fraco.

Há alguma complexidade no que acontece entre os valores extremos mas estamos interessado principalmente na ubicao de um entrevistado é entre um extremo e outro .

Em particular, podem ser definidos níveis qualitativos entre os extremos - estes são importantes e úteis na interpretação -

Este ponto ainda é uma idéia latente antes que algo claro. Embora os níveis qualitativos são definíveis , presume-se que os entrevistados podem estar em qualquer lugar no continuum do construto subjacente

Em resumo um mapa do construto pode ser considerado uma variável latente unidimensional .

Muitos construtos são mais complexos do que isto, por exemplo, pode ser multidimensional, mas isso não é uma barreira para a modelagem que fazemos, porque cada uma dessas dimensões pode ser considerado unidimensional e, portanto, podemos ter um mapa de construto para cada um deles dimensões.

Matriz de referencia

É o consenso do que e quanto deve conhecer o docente de ensino médio ao respeito de sua especialidade.

Este consenso é representado numa tabela de especificações ou matriz de referencia a qual geralmente é de dupla entrada onde em geral temos nas linhas conteúdos e nas colunas níveis cognitivos para estabelecer os pesos dos mesmos

Quadro 2. Exemplo de Matriz de referencia para uma prova educacional cognitiva

	Níveis Cognitivos			
	Nivel 1	Nivel 2	Nivel 3	Total
Conteúdo 1				
Conteúdo 2				
Conteúdo 3				
Conteúdo 4				
Conteúdo 5				
Total				

Um exemplo: Avaliação docente

Uma parte importante na formação de professores é estabelecer seu nível de desempenho ou de domínio nos principais conteúdos de ensino do currículo de sua especialidade, identificando estes segundo níveis cognitivos.

Em relação aos aspectos cognitivos vários esquemas ou quadros de referências têm sido propostos para ter em conta na preparação de provas. A modo de exemplo consideramos três níveis gerais os quais são apropriadas para medir diretamente através de um exame escrito

Outros níveis podem ser medido de forma abrangente com formatos maiores a um exame escrito, como a entrevista, as questões de desenvolvimento e avaliação de registros e atividades.

Quadro 3. Complexidade das tarefas baseada em níveis cognitivos para avaliações de professores de Ensino Médio

Nivel	Nome	Descrição
I	Gestão de informação	É o conjunto de questões que avaliam a capacidade em matéria de gestão de conceitos, termos e símbolos relacionados com as competências desejáveis que cada professor deve desenvolver dentro de um determinado assunto. Isso corresponde a um nível primário de cognição ligada ao passivo e concreto . Ele inclui reconhecimentos , descrições , sistemas, interpretações literais
II	Gestão de processos	É o conjunto de questões que avaliam a capacidade em matéria de gestão e implementação de estratégias (relacionamentos através de conceitos, imagens e procedimentos) relacionados com as competências desejáveis que cada professor deve desenvolver dentro de um determinado assunto. Isso corresponde a um nível primário de cognição ligado ao operacional e concreto. Involve o nível I, mas implica a aplicação directa desse nível para situações familiares .
III	Reflexão	É o conjunto de questões que avaliam a capacidade relativa à resolução de situações-problema envolvendo reflexão relacionada com as competências desejáveis que cada professor deve desenvolver dentro de um determinado assunto. Este nível corresponde a um nível secundário da cognição , vinculada ao operacional e ao hipotético dedutivo. Ele inclui os níveis 1 e 2

Quadro 4. Lista de conteúdos e níveis cognitivos (NC) a serem avaliados em provas hipotéticas por áreas dos professores

Àreas	Lista de conteúdos avaliados	Conteúdos	NC avaliados	Número de NC
Matemática y Lógico matemática	Ecuaciones, Triángulos, Conjuntos, Divisibilidad, Cuatro operaciones, Productos Notables	6	I, II y III	3
Letras	Gramática, Normativa, Redacción, Literatura peruana y Razonamiento verbal	5	I, II y III	3
Ciencia y ambiente	Anatomía y fisiología humana, Botánica, Anatomía y fisiología animal, Reinos biológicos, Ecología, Materia	6	I, II y III	3
Biología	Citología, División celular, Anatomía y fisiología animal, Genética y fisiología celular, Anatomía y fisiología humana y Botánica	6	I, II y III	3
Química	Estructura atómica, Enlaces químicos, Nomenclatura inorgánica, Cálculos químicos y Química orgánica	5	I, II y III	3
Física	Cinemática, Trabajo y energía, Electrodinámica, Estática / dinámica y Electromagnetismo	5	I, II y III	3
Ciencias Sociales	Perú prehispánico, Antropogénesis, Feudalismo, Perú siglo XIX – XX, El universo y Elementos básicos de economía	6	I, II y III	3
Filosofía	Religión y filosofía oriental, Filosofía y religión en el esclavismo, Filosofía en el feudalismo y cristianismo, Renacimiento, Filosofía en el capitalismo y Disciplinas filosóficas	6	I, II y III	3
Psicología	Ramas de la psicología, Estados de la conciencia, Bases biológicas del psiquismo humano, Bases socioculturales del psiquismo humano, Personalidad y Desarrollo humano	6	I, II y III	3
Unidocencia (Inicial, I Ciclo y II Ciclo)	Teoría de conjuntos, Cuatro operaciones, Gramática, Comprensión lectora, El cuerpo humano y Las regiones del Perú / Actividades económicas	6	I, II y III	3
Inglés	Vocabulario, Lenguaje, Reading y Writing	4	I y II	2
Educación Física	Habilidades y destrezas, Psicomotricidad, Juego, Capacidades físicas, Aprendizaje motor y Deporte	6	I, II y III	3
Computación	Diseño Macromedia, Diseño gráfico y Ofimática	3	II	1
Arte	Conceptos básicos de Arte, Historia del Arte, Arte y Cultura y Función social del Arte	4	I, II y III	3

LENGUAJE Y LITERATURA

		NIVELES COGNITIVOS			PREGUNTAS POR CONTENIDO
	CONTENIDOS	**I**	**II**	**III**	
1	Gramática	2	2	3	**7**
2	Normativa	1	2	3	**6**
3	Redacción	0	0	1	**1**
4	Literatura peruana	1	2	2	**5**
5	Razonamiento verbal	2	3	6	**11**
	PREGUNTAS POR NC	**6**	**9**	**15**	**30**
	PESO DE NC	**20%**	**30%**	**50%**	**100%**

Quando um mapa do construto é postulado pela primeira vez, é muitas vezes menos desenvolvidas do que aqui é apresentado. A melhora do mapa é obtido por meio de vários processos a medida que o instrumento é desenvolvido.

Esses processos incluem:
a) Explicar o construto a outras pessoas usando o mapa do construto;
b) Criar itens que você acredita que levam ao entrevistado a responder os níveis do mapa de construto;
c) Testar esses itens com uma amostra de respondentes e
d) Analisar os dados resultantes para verificar se os resultados são consistentes com as suas intenções expressas pela mapa do construto.

3. ELABORAÇÃO DE ITENS OU PLANEJAMENTO DA MEDIDA

Em seguida, logo após de ter o mapa do construto o medidor deve pensar em alguma forma como este construto teórico pode se manifestar em uma situação do mundo real.

No início, não será mais do que um palpite, um contexto em que se acredita que o construto deve estar envolvido, de fato, aquele em que o construto deve desempenhar um papel decisivo nesta situação.

Ainda este palpite se tornará mais cristalizada e se tornará em certos padrões.

A relação entre os itens e o construto não é necessariamente da forma como esta foi descrita. Muitas vezes, os itens podem ser pensados primeiro e o construto pode ser mais tarde elucidado.

Um item também pode assumir muitas formas, tais como múltipla escolha e Likert (tipos de itens de escolha forçada). Há muitas variações sobre isto. O entrevistado também pode produzir uma resposta livre de uma forma tal como um teste, entrevista ou desempenho (concertos, experimento científico, desenho).

Os itens variam em conteúdo e modo : perguntas de entrevista normalmente têm uma ampla gama de muitos aspectos de um tópico ; questões ou tarefas de um desempenho cognitivo podem ser apresentados dependendo das respostas a alguns itens iniciais; questoes em uma pesquisa podem usar diferentes conjuntos de opções e algumas resposta pode ser forçado e de resposta livre

Na situação mostrada na Figura 1,4 o medidor assume que o entrevistado " tem " uma certa quantidade da construto e que seu valor no construto é a causa das respostas dos itens no instrumento que o medidor usa.

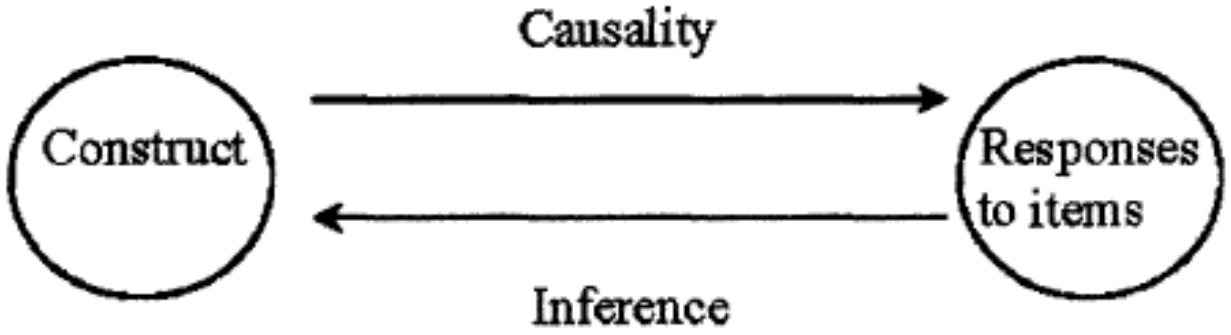


FIG. 1.4 A picture of the construct modeling idea of the relationship between degree of construct possessed and item responses.

No entanto, este agente causal é latente. O medidor não pode observar diretamente o construto. Em vez disso observa as respostas aos itens metros e então, inferir o construto subjacente a estas observações.

Note-se que a idéia de causalidade é uma suposição e a análise não fornece evidência dessa causalidade, na verdade, esta relação pode ser mais complexa.

Exemplo na Área de Língua

Nível I:

Qual é o substantivo relacionado à palavra abecedário?

a) Alfabeto b) Vogais c) Letras d) Maiúsculas e) Consonantes

Nível II:

Em qual dos itens abaixo estão escritos somente substantivos próprios?

a) Pedro, Augusto, Brasil, Bonito, Inteligente, América, Europa

b) Pedro, Augusto, Brasil, América, Europa

c) Pedro, Augusto, Brasil, Feio, Cachorro, América, Europa

d) Feio, Bonito, Cachorro, Inteligente

e) América, Europa, Bonito, Cachorro, Inteligente

Nivel III:
Associe os substantivos com seus respectivas classificações

I.	horário	a. primitivo
II.	caracol	b. derivado
III.	flotilha	c. composto
IV.	couve	d. coletivo

a) Ia, IIc, IIId, IVb b) Ib, IIa, IIId, IVc c) Ib, IIC, IIIa, IVd
d) Id, IIa, IIIb, IVc e) Ia, IIb, IIId, IVc

Atividades 1

Definir os conteúdos para avaliação de professores de uma determinada área de interesse (conteúdos mais relevantes usando programação anual)
5 minutos

Determinar a distribuição de questoes da matriz de referencia de sua área.
5 minutos.

Escrever um exemplo de item del área escolhida num nível cognitivo particular e conteúdo específico .
10 minutos

4. ANALISE CLASSICA DE ITENS OU MODELO DE MEDIÇÃO CLASSICO

Referências iniciais

- Gulliksen, H. (1950). Theory of Mental Tests. New York: John Wiley and Sons.
- Lord, F.M., Norvick, M.R. (1968). Statistical Theories of Mental Test Score. Reading: Addison-Wesley.
- Vianna, H.M. (1987). Testes em Educação. São Paulo: Ibrasa.

4.1. ANALISE ESTATÍSTICA DOS ITENS DE UM TESTE

- O objetivo geral em construção de testes é a obtenção de um teste de tamanho mínimo o qual deve produzir escores confiáveis e válidos para o uso que se desejar dar ao teste.
- 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 este analise nos permite garantir a validade e confiabilidade do instrumento que é construído.
- Os métodos que fazem parte da analise de itens dentro do enfoque da chamada Teoria Clássica dos Testes (TCT) se baseiam num conjunto de técnicas estadísticas de tipo descritivo as quais são interpretadas usando critérios empíricos e não supõem um modelo probabilístico.

Tamanho de 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.

Etapas da análise estatística dos itens

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

a) 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.

b) Qualificar os testes de acordo com a grelha de correção.

c) 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}
.	. . . . .
.Pn	X_{n1} 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.

Processo de Análise de Itens

Lazarte (op. cit) considera o seguinte processo numa análise de itens:

1. Decidir quais propriedades do escore total são importantes (ou seja, maximizar a variabilidade, maximizar a predição de critérios externos, etc.)
2. Identificar os parâmetros dos itens mais relevantes para estas propriedades do escore total.
3. Aplicar os itens para uma amostra de examinados que seja semelhante à população para a qual o teste está sendo construído.
4. Obter as estatísticas dos itens especificados no passo 2.
5. Estabelecer um plano para selecionar os itens ou identificar e revisar aqueles que estão com defeito.
6. Selecione um grupo final de itens.
7. Avaliar se o teste satisfaz o objetivo no passo 1, utilizando um estudo de validação cruzada.

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) Obter a distribuição de frequência dos escores totais e de cada subteste (se o teste tem subtestes).
b) Representar graficamente (polígonos de frequência ou histogramas) as distribuições de frequência dos escores totais e de cada subteste.
c) Calcular a média, variância, desvio padrão, assimetria e curtose da distribuição dos escores totais e parciais de cada subteste.
d) 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.

e) 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.
f) Calcular a dificuldade de cada item.
g) Calcular o poder discriminativo de cada item.
h) Calcular o coeficiente de validade de cada item.

Opcional:

i) Calcular a matriz de correlação entre os sub-testes, e entre o escore total e cada sub-teste.
j) 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.
k) Calcular a análise fatorial da matriz de intercorrelação dos itens para estabelecer a existência de fatores comuns.

4.2. TIPO DE ANALISES DOS ITENS

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:

A. Dificuldade do Item, média e variância

Itens dicotômicos.

São os mais comuns nos testes de aptidão. O item média 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:

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

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

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 Politômicos

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

B. 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:

$$D_i = p_{iS} - p_{il}$$

onde p_{iS} é a proporção de indivíduos no grupo superior que respondeu o item corretamente, e p_{il} 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:

Disc. = % de resposta correta no item i, do grupo Superior – % de resposta correta no item i 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 ou que favorece ao grupo inferior.

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.).

C. Validade do item.

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

c1) 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:

a) Correlação r de Pearson:

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

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

e 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.

b) Correlação bisserial:

É 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.

Versão corrigida

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 $\mathbf{p}_{\text{pbis}}$ é a correlação bisserial original entre o item e o escore total do critério, s_x é o desvio padrão do item, e $\mathbf{p}_{\text{ibis c}}$ 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.

c) <u>Coeficiente Phi:</u>

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, $\mathbf{p}i$, e a proporção de indivíduos que passam o critério, $\mathbf{p}_x$, é possível mostrar que o coeficiente Phi pode ser expressado como:

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

Onde $P_{i\,X}$ 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.

d) Índices de Confiabilidade e outros índices de 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

$$\sigma_i \rho_{i\,Y}\ ,$$

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:

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

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.

Quando o valor alfa é mais perto de 1 é melhor e significa que a soma das covariâncias dos itens em relação a variabilidade total é alta, indicando que os itens são consistentes entre si. Este índice também é chamado de consistência interna.

Cronbach's alpha	Internal consistency
α ≥ 0.9	Excellent (High-Stakes testing)
0.7 ≤ α < 0.9	Good (Low-Stakes testing)
0.6 ≤ α < 0.7	Acceptable
0.5 ≤ α < 0.6	Poor
α < 0.5	Unacceptable

https://en.wikipedia.org/wiki/Cronbach%27s_alpha

e) 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 analise 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 analise dos itens. Logo, no outro grupo, analisasse o escore total do teste baseado no itens selecionado na analise dos itens do primeiro grupo.
Quando as analises 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.

5. CRITERIOS PARA INTERPRETACAO DE RESULTADOS OU DO ESPAÇO DE RESULTADOS

Critérios para a escolha de itens

Finalmente quando todos os analises estatísticos dos itens são completados precisasse de uma revisão critica dos mesmos. Esta revisão deve ser feita considerando:

a) Analisar a dificuldade de cada um dos itens de modo que posa se formar grupos de dificuldade e fazer uma ordem entre eles.
b) Analisar a discriminação dos itens e retirar aqueles que tenham valores muito baixo, inferiores ao critério empírico recomendado.
c) Analisar a validez dos itens, removendo aquele que não satisfazem o critério mínimo considerado.
d) 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 analises 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 numero de itens fica pequeno, então precisasse fazer itens adicionais e aplicar a uma nova amostra e volver a fazer os analises apresentados.

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

6. ANALISIS DE ITENS USANDO SOFTWARE

Objetivo
Utilizando os códigos em SPSS e R abaixo, faça uma análise de Itens para os dados do Teste de Matemática para alunos de 6ta séria, e dados da Escala de Atitudes frente a Estatística de professores.

1. Código de Analise de itens usando SPSS

Dados de Conhecimentos
Use o seguinte códigos para analisar os dados mathb usando SPSS.

```
**  Chama os dados dicotomicos

GET
  FILE='C:\Users\Jorge Luis\Dropbox\Eventos\SBPC\Dia2\mathb.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
```

```
DATASET ACTIVATE DataSet1.
RELIABILITY
  /VARIABLES=i01 i02 i03 i04 i05 i06 i07 i08 i09 i10 i11 i12 i13 i14
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /STATISTICS=DESCRIPTIVE SCALE CORR
  /SUMMARY=TOTAL MEANS VARIANCE CORR.
```

Dados de Atitudes

Use o seguinte códigos para analisar os dados baseunionfinal usando SPSS. Note que o código salva as analises e ainda exporta os resultados para um arquivo Word.

```
** Chama os dados politomicos

GET
  FILE='C:\Users\Jorge
Luis\Dropbox\Eventos\SBPC\Dia2\baseunionoriginal.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
```

```
*Calcula o analise de items

DATASET ACTIVATE DataSet1.
RELIABILITY
  /VARIABLES=pre1 pre2 pre3 pre4 pre5 pre6 pre7 pre8 pre9 pre10 pre11
pre12 pre13 pre14 pre15 pre16
    pre17 pre18 pre19 pre20 pre21 pre22 pre23 pre24 pre25
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /SUMMARY=TOTAL.

*Organiza os proximos resultados segundo Pais

SORT CASES  BY pais$.
SPLIT FILE SEPARATE BY pais$.

* Calcula novamente a Analise de items por pais

DATASET ACTIVATE DataSet1.
RELIABILITY
  /VARIABLES=pre1 pre2 pre3 pre4 pre5 pre6 pre7 pre8 pre9 pre10 pre11
pre12 pre13 pre14 pre15 pre16
    pre17 pre18 pre19 pre20 pre21 pre22 pre23 pre24 pre25
```

```
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /SUMMARY=TOTAL.

* Conclue os reportes por pais

SPLIT FILE OFF.

* Salva as resultados

OUTPUT SAVE NAME=Document1
 OUTFILE='C:\Users\Jorge
Luis\Dropbox\Eventos\SBPC\Dia2\AnaliseitemsAtitudes.spv'
 LOCK=NO.
```

* Exporta os resultados em Word

```
* Export Output.
OUTPUT EXPORT
  /CONTENTS  EXPORT=ALL  LAYERS=PRINTSETTING  MODELVIEWS=PRINTSETTING
  /DOC                                        DOCUMENTFILE='C:\Users\Jorge
Luis\Dropbox\Eventos\SBPC\Dia2\AnalisedeItemsAtitudes.doc'
     NOTESCAPTIONS=YES  WIDETABLES=WRAP
```

```
     PAGESIZE=INCHES(8.266535433070866,                 11.69015748031496)
TOPMARGIN=INCHES(1.0)
    BOTTOMMARGIN=INCHES(1.0)
     LEFTMARGIN=INCHES(1.0)       RIGHTMARGIN=INCHES(0.9999999999999991).
LEFTMARGIN=INCHES(1.0)  RIGHTMARGIN=INCHES(0.9999999999999991).
```

2. Código de Analise de itens usando R

Dados de Conhecimentos

Use o seguinte códigos para analisar os dados mathb usando R. Instale previamente os pacotes que são indicados.

```
#analise de items dicotómicos
require(foreign)
pasta="C:\\Users\\Jorge Luis\\Dropbox\\ICMC\\SME0876\\Aula3SME0876"
setwd(pasta)
mathb=read.spss("mathb.sav")
a=data.frame(mathb)
mathbitems=a[,2:15]
```

```
require(psych)
alpha(mathbitems)
help(alpha)

require(epicalc)
alpha(vars=c(i01:i14), mathbitems)


require(psychometric)
item.exam(mathbitems, y=mathb$puntaje, discrim=TRUE)
help(item.exam)

#intervalo de Confiança Bootstrap para alpha
require(ltm)
cronbach.alpha(mathbitems, CI=TRUE, B=500)

#Correlaçao ponto biserial

biserial.cor(rowSums(mathbitems), mathbitems[[1]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[2]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[3]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[4]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[5]], level = 2)
```

```
biserial.cor(rowSums(mathbitems), mathbitems[[6]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[7]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[8]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[9]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[10]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[11]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[12]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[13]], level = 2)
biserial.cor(rowSums(mathbitems), mathbitems[[14]], level = 2)

require(coefficientalpha)
alpha.mathb=cronbach(mathbitems)
plot(alpha.mathb,type="d")
summary(alpha.mathb)
```

Dados de Atitudes

Use o seguinte códigos para analisar os dados baseunionfinal usando R.

```
#analise de items politomicos
################################

atitudes=read.spss("baseunionfinal.sav")
a=data.frame(atitudes)
atitudesitems=a[,4:25]
alpha(atitudesitems)

#analises de items por paises
atitudesP=subset(a,pais.=="Perú")
atitudesP
atitudesitemsP=atitudesP[,4:25]
alpha(atitudesitemsP)

atitudesE=subset(a,pais.=="España")
atitudesE
atitudesitemsE=atitudesE[,4:25]
alpha(atitudesitemsE)
```

3. Atividade 2

- Salve os resultados encontrados usando SPSS e R numa folha Excel para os dados dicotômicos e politômicos.
- Usando os critérios mostrados acima, compare os resultados obtidos por ambos programas. Indique quais resultados são obtidos em ambos programas, quais são diferentes.
- Compare seus resultados com os obtidos em IRTPRO

4. Alcances finais

As estatísticas mais apropriadas apresentadas pelo módulo SPSS e os diferentes softwares como R e IRTPRO 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.

Estes análises estatística dos itens sob perspectiva clássica, especialmente com o cálculo da média, variância e alfa de Cronbach se o item é eliminado, e a correlação item-total corrigida são os mais comuns e são válidos para o caso dicotômico e politômico. Más eles são analise básicas.

Referencias

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

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

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

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

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

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

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