Procedimentos para Estabelecer a Equivalência de Provas com Modelos da Resposta ao Item

Philip R. Fletcher*

RESUMO

Freqüentemente, nas avaliações de maior porte, as circunstâncias exigem formas alternativas da mesma prova ou provas de dificuldade variável. No primeiro caso, as formas são elaboradas para o mesmo nível de dificuldade mas variam por razões imprevisíveis. No segundo, as formas distintas são elaboradas para variar tanto em conteúdo quanto em dificuldade, muitas vezes por uma margem bastante expressiva. Nessas circunstâncias, o emprego de um método para estabelecer a comparabilidade dos resultados das múltiplas formas de prova se torna imprescindível. Recentes avanços em psicometria facilitam a tarefa de colocar todos os resultados na mesma escala, baseados na propriedade da invariância dos parâmetros de item obtida com a aplicação dos modelos de resposta ao item.

Philip R. Fletcher

PhD em Educação pela Universidade de Stanford, California - USA; Pesquisador Sênior da Fundação Carlos Chagas.

Existem diversas técnicas para comparar os resultados de provas distintas. Os procedimentos que estabelecem a *equivalência* das provas requerem formas alternativas da mesma prova. Pode-se estabelecer a equivalência das formas quando se referem ao mesmo conteúdo, têm a mesma fidedignidade e o mesmo nível de dificuldade. Nesse contexto, procura-se produzir formas alternativas da mesma prova, que sejam *idênticas*. O objetivo seria produzir múltiplas formas de uma prova que mede a mesma habilidade (característica, atributo, desempenho ou competência), destinadas à aplicação na mesma população de examinandos. As formas alternativas contêm perguntas diferentes mas são elaboradas, na medida do possível, para manter a mesma dificuldade, conteúdo e número de perguntas.

No entanto, apesar de todo o esforço realizado, geralmente não é possível produzir múltiplas formas com *exatamente* o mesmo nível de dificuldade. Quando se precisa avaliar o desempenho relativo de alunos que respondem a diferentes formas da mesma prova, o emprego de um método para estabelecer a equivalência dos resultados das múltiplas formas se torna imprescindível. Uma vez que se estabeleça a equivalência das formas, elas se tornam substituíveis entre si e será indiferente à qual das diversas formas o

(*) Filiação Institucional: Fundação Carlos Chagas

examinando responde.

Angoff (1978) delimita o tema da equivalência em provas ao apresentar quatro restrições ou exigências:

- as provas devem medir a mesma característica ou habilidade;

- a equivalência estabelecida deve ser independente dos dados particularmente usados para estabelecer esse princípio e deve ser aplicável em todas as situações parecidas;

- os escores de duas provas, uma vez estabelecida sua equivalência, devem ser substituíveis entre si;

- a equivalência deve ser simétrica ou a mesma, sem distinção da prova particularmente escolhida para ser a base.

O autor anota que a conversibilidade pode ser considerada apenas em situações onde as provas são *paralelas*, citando a definição de paralelismo por *Gulliksen (1950)*:

> Duas provas podem ser consideradas formas paralelas quando, após sua conversão para a mesma escala, suas médias, desvios padrão e correlações com todo e qualquer outro critério sejam iguais.

Por último, o autor enfatiza que apenas os escores de provas com coeficientes de erro iguais (fidedignidades iguais) devem ser considerados substituíveis.

Explica-se, por meio de uma analogia, por que não se pode estabelecer a equivalência entre provas de fidedignidade divergentes. Suponha que um treinador de futebol pretende selecionar jogadores que acertam pênaltis em pelo menos 75% das vezes. O treinador aplica uma forma reduzida da prova ao oferecer ao primeiro jogador quatro tentativas para acertar os pênaltis e aplica uma forma comprida da prova ao oferecer ao segundo jogador vinte tentativas para acertar. Supondo ainda que ambos os jogadores tenham uma precisão verdadeira de acertar 50% dos pênaltis, o primeiro jogador terá uma chance de 31% para alcançar o nível dos 75% de acertos em apenas quatro tentativas, enquanto o segundo terá uma chance de 2% para alcançar o mesmo nível em vinte tentativas. Isto ocorre porque quatro tentativas proporcionam bem menos fidedignidade do que vinte tentativas. Seria um equívoco tratar os resultados dessas duas provas sem levar em conta as diferenças de fidedignidade.

O tema da fidedignidade se desdobra com a questão da dificuldade da prova. Normalmente, os examinandos menos hábeis preferem provas de baixa fidedignidade que possam não revelar o verdadeiro nível de desempenho, enquanto os mais hábeis preferem provas de alta fidedignidade que revelem seu verdadeiro nível. Desconsideradas outras características das provas (número de itens, etc), uma prova de dificuldade apropriada para os examinandos é mais fidedigna do que uma prova de dificuldade imprópria (provas fáceis ou difíceis demais). Percebe-se, então, que o examinando não será indiferente quanto ao nível de dificuldade da prova e que provas de dificuldades divergentes não são substituíveis entre si, por causa das questões de fidedignidade.

As questões levantadas por *Angoff* foram reconsideradas e estendidas por *Lord (1977; 1980)*, quando defendeu as vantagens teóricas dos modelos da resposta ao item em procedimentos que estabelecem a equivalência de resultados de provas, A noção de escores substituíveis ganha uma nova dimensão com a introdução do conceito de eqüidade:

> Os escores transformados y^* e observados x podem ser considerados "equivalentes" quando, e apenas quando, é indiferente para cada examinando responder à prova X ou à prova Y *(Lord, 1977, p.128)*.

De acordo com esse princípio:

- Torna-se inviável a tentativa de estabelecer a equivalência de provas que medem diferentes características ou habilidades (consistente com a primeira restrição de *Angoff*);

- A equivalência de escores com margens de erro desiguais não pode ser estabelecida (consistente com os escores substituíveis);

e o mais extraordinário:

- Não se pode estabelecer a equivalência de provas que refletem diferentes níveis de dificuldade.

Nas próprias palavras de *Lord (1977, p. 128)*:

Se as provas X e Y têm dificuldades diferentes, a

relação entre seus escores verdadeiros é necessariamente não-linear, devido aos efeitos de piso e de teto. Se duas provas têm uma relação não-linear, é implausível que eles sejam igualmente fidedignos em todos os subgrupos de examinandos. Isto leva à conclusão incômoda de que, a rigor, não se pode tornar equivalentes os escores observados em provas de dificuldade diferente.

Lord demonstra que o requisito da eqüidade é satisfeito apenas no caso das provas infalíveis de fidedignidade absoluta. A equivalência dos escores observados poderia ser alcançado apenas no caso de duas provas realmente paralelas, ou seja, no caso trivial onde as duas provas são idênticas, item por item. Esse argumento, junto com o segundo requisito de Angoff, alimenta o interesse nos modelos da resposta ao item para estabelecer a equivalência das provas.

Os procedimentos de *calibração* são usados quando duas provas medem o mesmo conceito mas uma prova é mais comprida, mais fidedigna ou mais difícil que outra. Nessas circunstâncias, a equivalência das duas provas não pode ser estabelecida, mas elas podem ser calibradas na mesma escala. Normalmente, as técnicas de calibração são usadas para estabelecer uma única escala comum que permite comparar resultados de provas elaboradas para avaliar níveis de desempenho *distintos*. Exemplos típicos incluem baterias de provas comerciais disponíveis em diversos países, elaboradas propositadamente para aplicação nas diversas séries do ensino. Geralmente, as provas são aplicadas de acordo com o nível da série e com o propósito de acompanhar o desempenho dos alunos ao longo do ensino fundamental e secundário, ano a ano e série a série. A avaliação de alunos e de programas requer que os resultados dessas provas sejam comparados, o que obriga sua calibração numa única escala curricular.

Por exemplo: um aluno de 3ª série responde a uma prova elaborada para o nível da 3ª série. Na série seguinte, o mesmo aluno responde a uma prova do nível da 4ª série. Surge então a dúvida: como comparar os resultados das duas provas? Não é apenas porque os itens das duas provas sejam diferentes, mas porque as provas têm dificuldades diferentes, também. Os resultados das provas, em termos do número ou porcentagem de acertos em formas de diferentes níveis de dificuldade, não podem ser comparados diretamente. A solução é colocar todos os resultados numa única escala comum, onde resultados de múltiplas formas de uma prova, elaborada para medir a mesma característica em diferentes níveis de habilidade, podem ser comparados.

Quando múltiplas formas avaliam o conteúdo no mesmo nível de dificuldade, a literatura e a prática referem-se ao processo de estabelecer a *equivalência horizontal* das provas *(horizontal equating)*. A perseguição desse objetivo gera a maior parte do interesse na questão da equivalência, origem dos métodos que garantem a produção contínua e sustentada de formas alternativas e substituíveis de provas que produzem resultados idênticos ano a ano. Quando múltiplas formas avaliam o conteúdo em diferentes níveis de dificuldade, a literatura faz referência ao processo de estabelecer a *equivalência vertical* das provas *(vertical equating)*. Tanto do ponto de vista teórico quanto prático, esta é considerada uma questão muito mais complexa que a equivalência horizontal. O desacordo geral dos teóricos em medidas sobre esse tema constitui a razão pela qual se considera a equivalência vertical uma questão muito mais intrigante e desafiante que a equivalência horizontal. Em jogo está nada menos do que a nossa capacidade de medir o crescimento cognitivo.

Apesar do uso bastante generalizado do termo "equivalência vertical", é recomendável restringir o uso do conceito da equivalência a situações onde os escores de formas alternativas de uma prova podem ser usados indistintamente. Nesse contexto, o adjetivo "horizontal" se torna redundante e desnecessário. Devido ao fato de que os escores de escala em diferentes níveis de desempenho não podem ser usados indistintamente, um termo sem conotações de equivalência se aplica melhor no segundo contexto. Parece plausível pressupor que o conteúdo apresentado num conjunto de provas elaboradas para uso em diversos níveis de ensino representa um *contínuo desenvolvimento* para determinado currículo. As técnicas de construção de *escala* associam números ao desempenho de examinandos para representar uma seqüência ordenada de níveis de uma característica ou habilidade. O resultado dessa atividade é uma *escala de esco-*

res, e uma escala desenhada para representar esse contínuo de desenvolvimento seria então uma *escala curricular*. Portanto, parece mais correto dizer que as diversas formas da prova foram *calibradas* numa escala comum ou escala comparável e que os escores que resultam dessa transformação são *escores calibrados* ou *escores comparáveis*. A preferência pelo termo "calibração" é nitidamente associada à popularização dos modelos da resposta ao item durante a década de 80.

Ao estabelecer a equivalência de formas substituíveis ou a calibração de formas distintas, o objetivo é sempre de obter escores *na mesma escala*. No caso da equivalência, as formas são elaboradas para o mesmo nível de dificuldade mas variam por razões imprevisíveis, enquanto, no contexto da calibração, as formas distintas são elaboradas para variar tanto em conteúdo quanto em dificuldade, muitas vezes por uma margem bastante expressiva.

Nem por isso esgotam-se todas as possibilidades de relacionar resultados de provas. Técnicas de *projeção*, geralmente por análise de regressão, são apropriadas em circunstâncias onde nenhum dos requisitos da equivalência se aplica. Com as técnicas de projeção, duas provas não precisam de fidedignidade/dificuldade igual e nem precisam medir a mesma característica ou habilidade. O objetivo de projeção é prever ou predizer os escores de uma prova a partir dos escores de outra prova. Tanto a calibração quanto a projeção estabelecem uma relação entre duas ou mais provas. No caso da calibração essa relação se estabelece a partir de uma escala comum, enquanto a projeção estabelece a relação por via de uma fórmula de predição. Tanto as técnicas de calibração quanto as de predição exigem mais trabalho para assegurar que os resultados das análises são válidos. A contribuição do erro de transformação aos escores transformados pode não ser igual ao longo da amplitude da escala. Geralmente os escores transformados têm maiores erros nas caudas das distribuições. Os procedimentos precisam ser verificados para assegurar que se apliquem nos diversos subgrupos da população.

Existe considerável interesse nos modelos da resposta ao item, pelas vantagens teóricas e práticas que prometem em situações que requerem escores comparáveis. Apresentamos a seguir uma introdução à aplicação desses modelos em procedimentos que estabelecem a equivalência de múltiplas formas de uma prova.

Uma das propriedades básicas que sustenta o uso dos modelos da resposta ao item, em procedimentos para obter escores na mesma escala, é que os escores produzidos para representar o desempenho do examinando não dependem da dificuldade da prova. Quando os dados da prova satisfazem os requisitos do modelo da resposta ao item, obtém-se uma estimativa do parâmetro de habilidade do examinando independente do conjunto de itens (forma da prova) respondido. Satisfeitos os mesmos requisitos, obtêm-se estimativas dos parâmetros de item, independentes do grupo particularmente examinado. Onde se obtêm estimativas de item livres da amostra, obtêm-se também estimativas de habilidades livres da prova. Portanto, quando um examinando participa de uma prova, não lhe importa o grau de dificuldade da forma particular que recebe; sua estimativa de habilidade obtida a partir de qualquer forma da prova será sempre a mesma, à parte o erro da medida, uma vez que as estimativas dos parâmetros de item foram colocadas na mesma escala. Desaparece a preocupação com a dificuldade variável das diversas formas da prova. O único problema que requer atenção é lembrar que os itens calibrados independentemente precisam ser colocados na mesma escala, antes de gerar parâmetros de habilidade para os examinandos.

Os modelos de resposta geram verdadeiras escalas de intervalo. A escala de intervalo preserva a relação de ordem entre elementos encontrados no mundo real e ainda postula a existência de uma correspondência exata entre esses elementos no âmbito comportamental e os números reais da escala, considerados arbitrários apenas o ponto de origem da escala e a unidade de medida de sua métrica. Atribui-se assim significado não apenas aos valores da escala de habilidade e a sua ordem relativa, mas também às diferenças relativas entre valores encontrados ao longo dessa escala. A diferença de unidade medida, que representa o intervalo entre dois valores, tem o mesmo significado para a propriedade representada no limite inferior da escala que um intervalo de igual tamanho encontrado

em seu limite superior. Essa propriedade das escalas de intervalo dos modelos de resposta permite que se compare a mudança no desempenho de duas pessoas de habilidade desigual. Isto permite a análise de escores de diferença em pesquisas longitudinais.

Devido a uma relação matemática exata entre estimativas de habilidades e escores verdadeiros estimados nos modelos de resposta ao item, podem-se aproveitar os escores de habilidade para estabelecer a relação entre escores verdadeiros estimados em duas formas de uma prova. Ao estabelecer essa relação, os parâmetros de item e de habilidade se tornam equivalentes mediante uma simples transformação linear. Após essa transformação, os parâmetros de habilidade podem ser comparados diretamente; não há nenhuma necessidade de estabelecer a comparabilidade das escalas do número de acertos para diferentes formas de prova. Os três requisitos da *eqüidade, invariância e simetria* se satisfazem quando os escores verdadeiros são usados. Portanto, é viável tornarem-se equivalentes formas alternativas de dificuldade divergente expressiva, quando se usam os escores verdadeiros dos modelos da resposta ao item. Além do mais, os examinandos usados para estabelecer a equivalência não precisam vir de amostras aleatórias da mesma população para determinar uma relação de equivalência que seja invariante em diversos grupos. Devido a sua importância central nos procedimentos de estabelecer a equivalência de provas de dificuldade desigual, examinamos a seguir algumas características dos modelos da resposta ao item.

De acordo com a teoria das características latentes, o desempenho de um examinando numa prova pode ser explicado por certas características não observáveis do examinando, suas *características latentes* ou *habilidades subjacentes.* O objetivo é caracterizar o examinando em termos dessas características e usar essas informações para prever seu desempenho futuro em outras provas e tarefas afins. Modelos matemáticos são incorporados a essa teoria para definir uma relação explícita entre as habilidades subjacentes, não observáveis, e o desempenho observável das respostas de prova.

Como primeira aproximação, pode-se pressupor que *várias* características latentes afetam o desempenho do examinando num item de prova. Essas características constituem as dimensões psicológicas necessárias para uma descrição psicológica do examinando *(Lord & Novick, 1968, p. 359).* A descrição completa é representada pelo vetor:

$$\theta = (\theta_1, \theta_2, \ldots \theta_k),$$

onde cada característica é representada pela letra grega θ. O conjunto de valores nesse vetor posiciona o examinando num espaço latente de *k* dimensões. Pressupondo que todas as características latentes para uma resposta ao item estejam presentes no vetor, então, a distribuição condicional dos escore de item seria idêntica para θ fixo em todas as populações.

A regressão do escore de item em θ é chamada a *função de característica do item.* Devido à abrangência da definição do espaço latente, a função de característica do item permanece necessariamente invariante em todos os grupos de examinandos. Segue, então, que qualquer parâmetro que define a função da característica do item é um parâmetro de item *invariante.* A situação é análoga à regressão múltipla quando todas as variáveis relevantes estejam presentes na análise. Nesse caso, os coeficientes das variáveis analisadas se tornam estáveis de uma análise para outra, desconsiderado o erro amostral.

Ao notar que a maioria das provas são elaboradas para avaliar o desempenho em apenas *uma* área ou matéria, na maioria dos casos, parece plausível pressupor que o espaço latente de uma prova tenha apenas uma dimensão. Com $k=1$, pressupõe-se que a prova esteja composta de itens suficientemente homogêneos que medem apenas uma única característica ou habilidade latente. Nestas circunstâncias, o modelo é unidimensional e a função de resposta ao item é chamada a *curva de resposta ao item.* Ao afirmar que todas as características latentes estejam presentes nesse modelo unidimensional, as respostas para examinandos com o mesmo nível de habilidade são independentes entre si *(Lord & Novick, 1968, p. 361).* A condição de independência é essencial para a estimação de parâmetros de item e de pessoa nos modelos de resposta. Quando os resultados da prova estão em conformidade com o modelo da resposta ao item, obtém-se a invariância dos parâmetros de

item e de habilidade. Para um conjunto de itens assim calibrados, a mesma estimativa de habilidade para o examinando é obtida em qualquer subconjunto de itens. Também, para qualquer subamostra de examinandos, os parâmetros de item são iguais. Isto tem uma enorme conse qüência para a tarefa de tornar os resultados de prova comparáveis entre si. Com conjuntos de item calibrados separadamente, as diferenças intencionais e imprevistas da dificuldade dos itens podem ser controladas e compensadas.

Uma grande variedade de modelos matemáticos da relação funcional que une as respostas de item à característica latente já foi experimentada. Em parte, o contencioso dos adeptos de cada modelo gira em torno do mérito relativo da parcimônia e elegância das formulações mais simples, de um lado, e o maior realismo e poder de predição dos modelos mais complexos, do outro. Evidentemente, essa não seria uma questão que pudesse ser resolvida lógica ou empiricamente, porque, de certa forma, os objetivos perseguidos representam alternativas distintas.

No entanto, esses modelos também representam perspectivas filosóficas radicalmente diferentes sobre a teoria das medidas. A maior parte da controvérsia gira em torno do uso do modelo *Rasch*, baseada numa função logística de apenas um parâmetro, e o modelo de três parâmetros defendido por *Lord*, entre outros. Os adeptos do modelo de três parâmetros notam que o acerto casual é sempre presente nos itens de múltipla escolha, o tipo de item usado quase que exclusivamente nos grandes empreendimentos. Do outro lado, aqueles que favorecem o uso do modelo *Rasch* alegam que o acerto casual é característica do examinando e não do item; em todo caso, o parâmetro que representa o acerto casual não pode ser estimado com confiança. O modelo *Rasch* recebe críticas severas por ignorar o acerto casual e por não levar em conta o poder discriminatório variável dos itens. Nossa apresentação se restringe ao modelo de três parâmetros.

Esse modelo pressupõe que a probabilidade de um acerto possa ser representada por uma função logística de três parâmetros. A probabilidade do acerto do item *j* varia de acordo com a habilidade latente do examinando θ, de acordo com a função:

$$P_j(\theta) = c_j + \frac{1 - c_j}{1 + e^{-Da_j(\theta - b_j)}}.$$

No gráfico 1, cada um dos três parâmetros de item, b_j, a_j e c_j tem uma nítida interpretação comportamental:

A Curva de Característica do Item:

A Probabilidade de um Acerto Pj (θ) em Diferentes Níveis de Habilidade θ

• O parâmetro c_j, representado pela linha reta horizontal no fundo do gráfico, define o limite inferior da probabilidade de um acerto, $P_j(\theta)$. O parâmetro representa a probabilidade de um *acerto casual* quando o examinando desconhece a resposta certa. Nos itens de resposta aberta, quando não houver possibilidade para um acerto casual, considera-se c_j=0.

• O parâmetro b_j, representado no gráfico pela linha reta vertical, desloca a curva de característica para a direita ou esquerda, dependendo do grau de dificuldade do item *j*. O b_j seria, portanto, o parâmetro da *dificuldade* do item. O valor de b_j assinala o ponto na escala de habilidade θ onde a ogiva da logística de $P_j(\theta)$ encontra sua inflexão. Devido à simetria da curva em torno desse ponto $P_j(b_j)$ é o nível de habilidade médio onde a probabilidade de um acerto alcance o ponto intermediário entre c_j e 1,0.

• O parâmetro a_j, representado no gráfico pela linha tangencial angular, é proporcional à inclinação da curva de $P_j(\theta)$ no ponto de sua inflexão e deve ser interpretado como o parâmetro do *poder discriminatório* do item. Tanto maior o valor de a e o grau de inclinação da curva de característica em torno de *b*, maior seria o poder discriminatório do item nesse ponto da escala. Para aumentar a informação obtida pela prova, procuram-se altos valores de *a* e uma boa distribuição dos valores de *b*, consistente com a distribuição de habilidade dos examinandos.

Por último, a constante $D = 1{,}7$ é usada para impor a escala de uma ogiva normal padrão. Com essa transformação e quando $c = 0$, o valor de *b* é igual à média e o valor de $1/a$ igual ao desvio padrão da distribuição cumulativa normal. Pode ser demonstrado: quando $D = 1{,}7$ e $c = 0$, a ogiva logística diverge da ogiva normal por menos de 1 ponto percentual para todos os valores de θ.

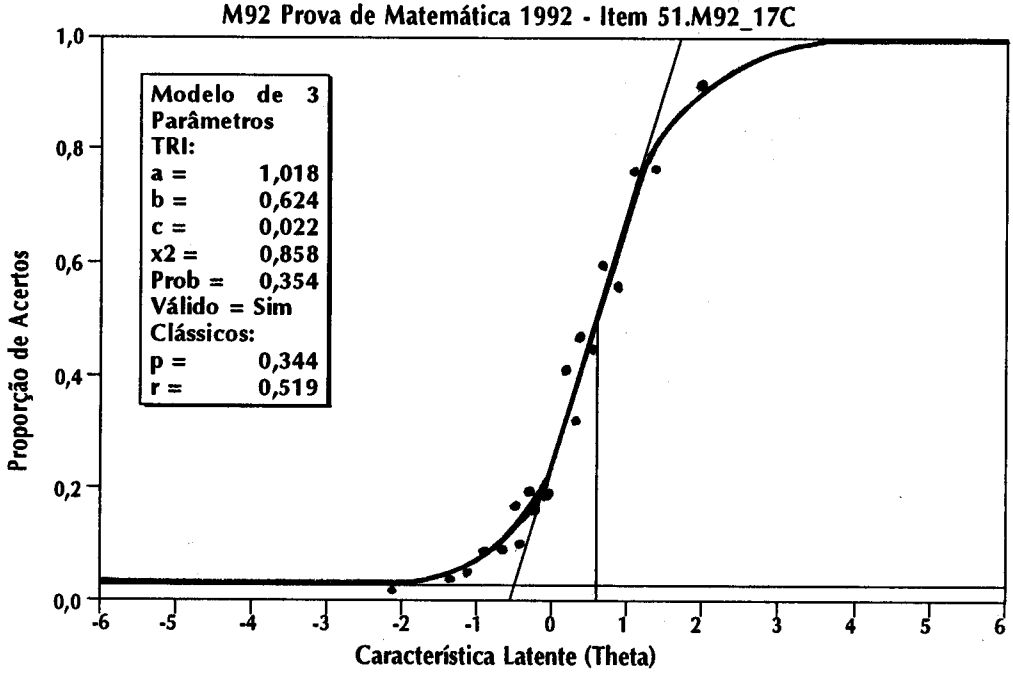

GRÁFICO 1

No gráfico, os pontos em torno da ogiva representam pessoas do grupo de examinandos reunidas segundo valores de habilidade. Os pontos mostram a porcentagem de acertos do item em cada grupo. A dispersão dos pontos em torno da curva representa o grau de ajuste da função logística aos dados observados da amostra.

Estimam-se os valores dos parâmetros do modelo de resposta aplicando técnicas de máxima verossimilhança. Procuram-se os valores para um conjunto de parâmetros que permite determinado conjunto de dados observados aparecer mais verossímil à luz do modelo matemático adotado. O relevante conjunto de dados é a configuração de respostas de um determinado examinando, o vetor $x_i = (x_1, x_2, ..., x_n)$, repetido para as *N* pessoas na amostra, o que forma uma matriz com linhas de respostas para *pessoas* e colunas de respostas para *itens*:

$$X_{ij} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_{N1} & x_{N2} & \cdots & x_{Nn} \end{bmatrix}$$

Quando existe independência condicional entre dois eventos com probabilidades distintas P(A) e P(B), a probabilidade da ocorrência simultânea dos dois eventos juntos é dada por:

$$P(A \cap B) = P(A) \,.\, P(B).$$

Para estimar o valor do parâmetro de habilidade para uma pessoa, pressupõe-se que todas as dimensões do problema estejam representadas no modelo, o que estabelece a independência entre as respostas de item. Nessas circunstâncias, a probabilidade do conjunto de respostas x_i do indivíduo *i* é dada por:

$$L_i(\beta) = \Pr\{x_i \mid \theta\} = \prod_{j=1}^{n} P_j(\theta)^{x_{ij}} Q_j(\theta)^{1-x_{ij}},$$

ou seja, o produto contínuo de $Pj(\theta)$ ou $Qj = 1 - Pj(\theta)$ para todos os *n* itens da prova, de acordo com o acerto ou erro da pessoa em sua resposta a cada item *j*. Quando se trata de uma afirmação teórica, o valor obtido acima representa a *probabilidade* da ocorrência do conjunto de resposta x_i condicional em θ, ou seja, $\Pr\{x_i \mid \theta\}$. Quando se calcula o mesmo valor usando esco-

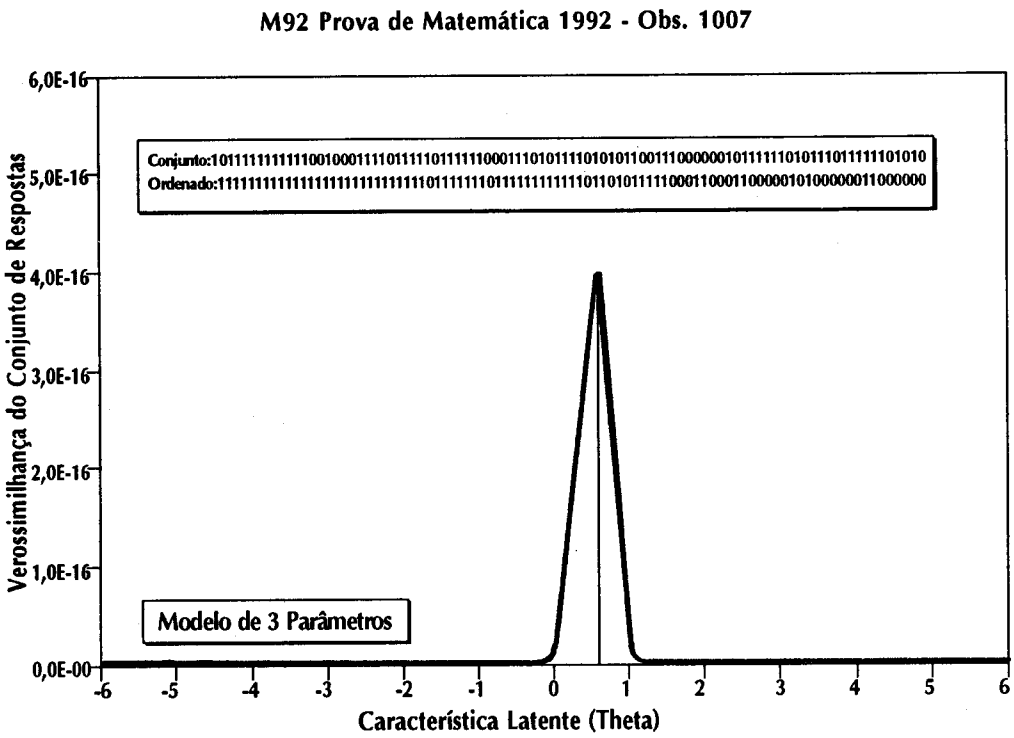

GRÁFICO 2

res de item observados, trata-se de uma função de *verossimilhança (likelihood)* do parâmetro de habilidade θ, ou seja, $L_i(\beta)$. A solução de *máxima verossimilhança* seria simplesmente a seleção do valor do parâmetro de habilidade θ que maximiza o valor de $L_i(\beta)$. Métodos numéricos são usados para encontrar esse máximo. O valor de θ correspondente ao máximo constitui a estimativa de máxima verossimilhança de θ_i (Gráfico 2).

Estimação do Parâmetro de Habilidade θ pela Técnica da Máxima Verossimilhança

Para estimar os três parâmetros de item, pressupõe-se novamente que todas as dimensões estejam representadas, o que estabelece a independência entre as respostas das pessoas na amostra. A probabilidade do conjunto de respostas para um item x_j é

$$L_i(\alpha) = \Pr\{x_j \mid a,b,c\} = \prod_{i=1}^{N} P_j(\theta)^{x_{ij}} Q_j(\theta)^{1-x_{ij}},$$

que define a probabilidade teórica de x_j condicional nos valores de *a, b, c,* isto é, $\Pr\{x_j \mid a,b,c\}$, ou a verossimilhança $L_i(\alpha)$, quando os parâmetros do modelo forem ajustados aos dados observados. Devido à independência dos itens, a tarefa de estimar simultaneamente todos os parâmetros de item é idêntica à estimação dos parâmetros de cada item em seqüência. Os procedimentos usados envolvem processos iterativos que refinam as estimativas preliminares dos parâmetros em etapas sucessivas, prosseguindo assim até que convirjam para valores estáveis.

O processo é formalmente equivalente à estimação dos parâmetros de uma função logística em biologia, onde a variável dependente é uma ocorrência qualquer codificada zero e um. Esse procedimento pressupõe que o valor do examinando na escala da característica latente seja um dado fixo. No entanto, em psicologia, a característica latente é uma construção teórica e não uma realidade física encontrada logo à mão. Surge, então, a dúvida sobre o que deveria ser usado para estimar o escore de escala na primeira

etapa de estimação. Uma das soluções propostas é substituir o desvio normal correspondente à proporção de itens acertados na prova, como uma estimativa preliminar da habilidade na primeira etapa.

Utilizando esses valores preliminares para os parâmetros de habilidade passa-se a estimar os parâmetros de item. Em seguida, pressupõe-se que os valores dos parâmetros de item estimados nessa etapa sejam os verdadeiros e calculam-se novas estimativas de habilidades que maximizam a função de verossimilhança. Repete-se esse ciclo iterativo de dois passos, até que as estimativas dos valores de parâmetro se estabilizem, permanecendo essencialmente iguais de uma etapa para outra. O procedimento é curioso porque a variável considerada dependente nessa relação (θ) não é fixa e muda em função das variáveis independentes a cada etapa de estimação.

Os procedimentos para estabelecer a equivalência de duas ou mais provas com modelos de resposta ao item envolvem três passos. Pressupondo que um modelo satisfatório já foi escolhido, o primeiro passo envolve a seleção de um plano para a coleta de dados que permita estabelecer a equivalência das provas. O segundo passo coloca as estimativas dos parâmetros de item na mesma escala, dentro da estrutura desse plano. O terceiro passo gera estimativas de habilidade comparáveis para os examinandos mediante técnicas de máxima verossimilhança.

Existem basicamente três planos na coleta de dados para se obter resultados comparáveis. O primeiro plano envolve um *grupo único*, onde as mesmas pessoas respondem às diversas formas da prova. Quando um único grupo responde a todas as formas da prova, as diferenças nas dificuldades dos itens não seriam comprometidas e distorcidas por diferenças de habilidade entre examinandos. Todos os itens são calibrados diretamente na mesma escala. O segundo plano envolve *grupos aleatórios* escolhidos da mesma população. Quando os grupos representam verdadeiras amostras probabilísticas da mesma população, a distribuição dos níveis de habilidade seriam aproximadamente iguais em cada grupo. À maneira do plano que envolve um grupo único, as diferenças nas dificuldades dos itens não se confundem com diferenças de habilidade entre grupos e os parâmetros de item podem ser calibrados diretamente na mesma escala. O terceiro plano envolve uma *prova âncora* administrada a dois ou mais grupos. Os resultados da prova âncora servem para controlar e compensar as diferenças nos níveis de habilidade dos grupos. A prova âncora pode ser uma prova única ou várias provas que contenham um núcleo de itens comuns. Esse último plano explora melhor as vantagens dos modelos de resposta ao item.

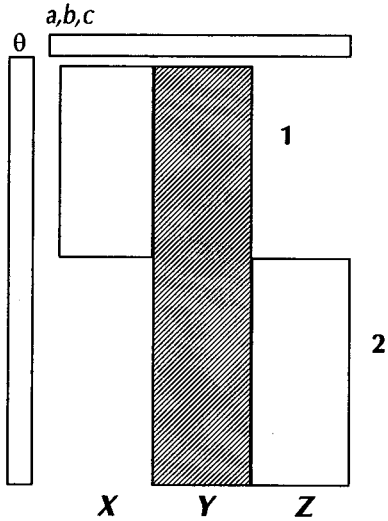

Figura 1

Método Conjunto para Estabelecer a Equivalência

No caso da prova âncora aplicada em dois grupos, o grupo 1 recebe os conjuntos de itens ***X*** e ***Y***, enquanto o grupo 2 recebe os conjuntos ***Y*** e ***Z***. Os itens no conjunto ***Y*** são os chamados itens comuns que servem como âncora, relacionando os dois grupos. Com os modelos da resposta ao item, podem-se estimar os parâmetros de item para todos os três conjuntos de item de uma só vez, utilizando os dados de ambos os grupos de examinandos (figura 1). Esse método conjunto coloca as estimativas dos parâmetros de item para os três conjuntos na mesma escala automaticamente, porque foram obtidos juntos, sujeitos apenas a quaisquer restrições impostas para definir a unidade e origem da escala de habilidade (θ).

O método tem a vantagem de não propagar erros de transformação quando envolve um grande número de provas. A desvantagem é que requer muita atenção para verificar se o modelo de resposta se aplica satisfatoriamente aos dois grupos. Quando essa condição não se satisfaz, prossegue-se à reestimação dos parâmetros, permitindo que os parâmetros de item flagrados antes se ajustem independentemente nos dois grupos.

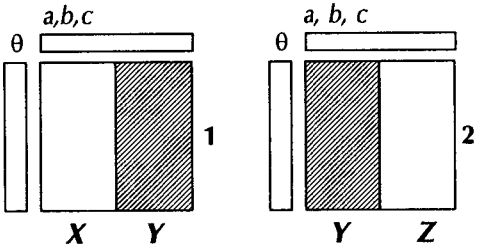

Figura 2

Método Paralelo para Estabelecer a Equivalência

Com o método paralelo, os valores dos parâmetros para os conjuntos de item ***X*** e ***Y*** são estimados com os dados do grupo 1. Em outra calibração, os valores dos parâmetros para os conjuntos de ***Y*** e ***Z*** são estimados com os dados do grupo 2. Dessa maneira, obtêm-se dois conjuntos de estimativas para os parâmetros de item do conjunto ***Y***, um conjunto baseado nos dados do grupo 1 e outro conjunto baseado nos dados do grupo 2. Estas estimativas dos parâmetros dos itens comuns são usadas para obter estimativas de coeficientes de transformação linear, para transformar a escala (θ) do grupo 2 na métrica da escala de (θ) do grupo 1. Os coeficientes de transformação são usados para transportar todos os parâmetros do grupo 2 para a escala de θ do grupo 1. O método paralelo é especialmente útil para iluminar os processos que estabelecem a equivalência.

Os modelos logísticos que representam a probabilidade de acerto $P_j(\theta)$ são funções da quantidade $a_j(\theta - b_j)$. Caso se acrescentasse a mesma constante β a cada θ e b_j, a quantidade $a_j(\theta - b_j)$ permaneceria inalterada e assim, também, a função de resposta ao item que representa $P_j(\theta)$. Da mesma maneira, se cada θ e b_j fosse multiplicado pela mesma constante α, as quantidades $a_j(\theta - b_j)$ e $P_j(\theta)$ permaneceriam inalteradas. Portanto, é claro que a *origem* e *unidade de medida* da escala de habilidade são arbitrárias. Podemos escolher *qualquer* escala para θ, desde que escolhamos a mesma escala para b_j. Essa *indeterminância da escala* de intervalos dos modelos de resposta acaba se tornando uma grande vantagem. Se bem que a origem e as unidades de medida sejam arbitrárias, duas medidas alternativas da mesma habilidade podem ser facilmente colocadas na mesma escala mediante uma simples transformação linear. Exploramos a propriedade da indeterminância de escala para obter uma escala curricular comum.

A indeterminância de escala significa que, se estimamos independentemente os parâmetros b_j e a_j de um conjunto de itens para dois grupos de examinandos ξ e δ, os dois conjuntos de parâmetros resultantes não seriam idênticos. No entanto, os dois conjuntos de θ e b_j devem ter uma relação linear de um para o outro. Ou seja, $b_\xi = \alpha b_\delta + \beta$ e $\theta_\xi = \alpha\theta_\delta + \beta$, desconsiderada a variabilidade amostral. Da mesma maneira, os dois conjuntos de a_j devem ser iguais, uma vez que se leve em conta a unidade de medida da escala. Ou seja, $a_\xi = a_\delta/\alpha$, ignorada a variabilidade amostral. Os dois conjuntos de c_j devem ser iguais, uma vez que os c_j aparecem no modelo na dimensão das probabilidades de resposta e não na dimensão de habilidade. Portanto, os coeficientes de acerto casual não sofrem nenhum efeito da mudança de origem ou de unidade de medida da escala de habilidade.

O procedimento mais comum para estabelecer a escala dos parâmetros para um único conjunto de itens é fixar a média e o desvio padrão de θ em 0 e 1, respectivamente, para cada grupo investigado. Quando se estimam parâmetros de item e de habilidade separadamente para cada prova, as escalas de parâmetro das duas provas terão uma relação linear e $\theta_\xi = \alpha\theta_\delta + \beta$. Uma vez que se determinam as constantes α e β, os parâmetros de item e de habilidade das duas provas se tornam comparáveis, ou seja, os resultados são apresentados numa *escala comum*.

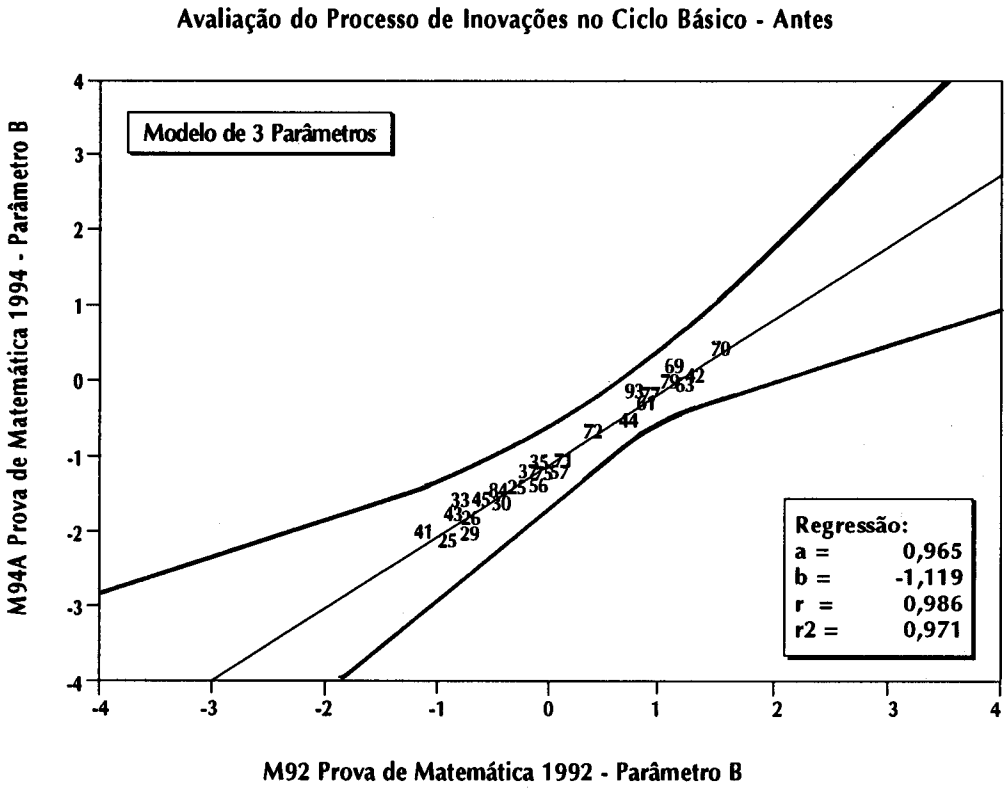

GRÁFICO 3

A Invariância dos Parâmetros de Dificuldade Usada para Estabelecer a Equivalência de Escala em Duas Provas

Existem várias alternativas para determinar as constantes α e β. Um método que tem um certo apelo intuitivo utiliza apenas os parâmetros de dificuldades dos itens, b_j (gráfico 3). Devido à propriedade da invariância dos parâmetros de item e à propriedade da indeterminância de escala, os mesmos itens, quando calibrados independentemente em dois grupos, devem se relacionar por uma simples transformação de origem e de escala. Isto significa que num gráfico dos parâmetros b_j, entre itens comuns a duas provas devem cair ao longo de uma *linha reta*. Os coeficientes da inclinação e de intercessão da regressão que definem essa reta seriam as constantes de transformação α e β. Com a aplicação dessas novas constantes aos parâmetros originais, obtêm-se novas estimativas para os parâmetros de dificuldades: $b_\xi = \alpha b_\delta + \beta$

parâmetros de discriminação: $a_\xi = a_\delta / \alpha$

e parâmetro de habilidade: $\theta_\xi = \alpha\theta_\delta + \beta$

O modelo da segunda prova é transportado integralmente para a métrica da primeira prova.

Após essa transformação, existe uma correspondência entre os parâmetros de dificuldade bj ao longo da escala de habilidade, em uma e outra prova. A nova identidade é representada por uma linha de regressão com intercessão de zero e inclinação de um. As dificuldades dos itens das duas provas se correspondem, item por item, ao longo da escala de habilidade. A partir desse ponto, as duas provas devem ser consideradas equivalentes.

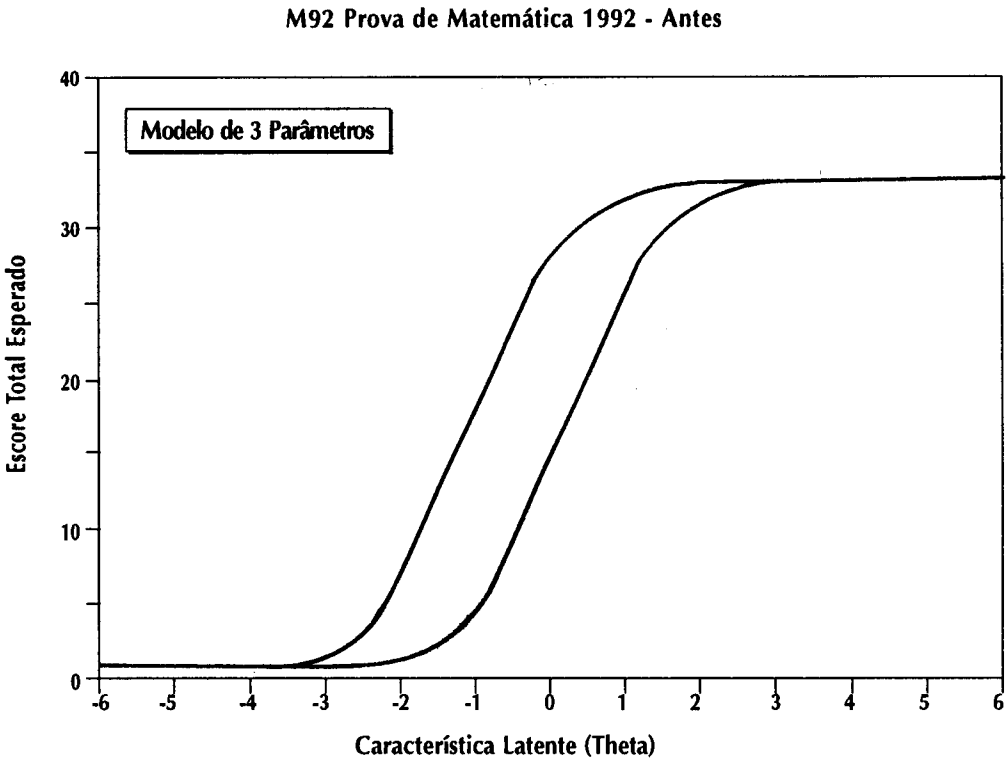

GRÁFICO 4A

Escores Verdadeiros em 33 Itens Comuns a Duas Provas Antes da Transformação Linear

O processo de transformação desenvolvido por *Stocking* e *Lord (1983)* tem a vantagem de considerar simultaneamente os parâmetros a_j e b_j. O procedimento envolve a transformação da *curva de característica da prova* (gráfico 4A). Se o escore esperado do examinando *i* no item *j* é $P_j(\theta_i)$, o número de acertos esperado numa prova de *n* itens seria $\sum^n_{j=1} P_j(\theta_i)$. Na teoria clássica, essa expectativa é chamada *escore verdadeiro do número de acertos*. O gráfico do número esperado de acertos em relação a (θ) revela a chamada curva de característica da prova. A transformação das curvas de característica proposta por *Stocking* e *Lord* minimiza a diferença entre os escores verdadeiros dos itens comuns. Supondo que as estimativas de item fossem livres de erro, a escolha dos parâmetros lineares levaria os escores esperados a coincidirem em cada valor de θ. Os coeficientes de transformação linear α e β, obtidos por essa transformação, são aplicados em seguida a todos os parâmetros do modelo.

Para compreender o processo, suponha que provas ξ e δ são ambas medidas da mesma habilidade θ e que os parâmetros de item para ambas as provas foram colocados numa escala comum para que $\theta_\xi = \alpha\theta_\delta = \theta$. Nesse caso, os escores verdadeiros do número de acertos ζ e η nas provas ξ e δ, respectivamente, relacionam-se com θ através de suas funções de característica de prova (a soma das funções de resposta ao item dos itens comuns às duas provas), isto é:

$$\zeta = \zeta(\theta) = \sum P_j(\theta) \quad \text{e}$$

$$\eta = \eta(\theta) = \sum P_j(\theta).$$

As variáveis ζ , η e θ são todas medidas da mesma característica ou habilidade. Elas se distinguem apenas pela escala numérica: as escalas de ζ e η dependem do número de itens comuns às duas provas, enquanto a escala da me-

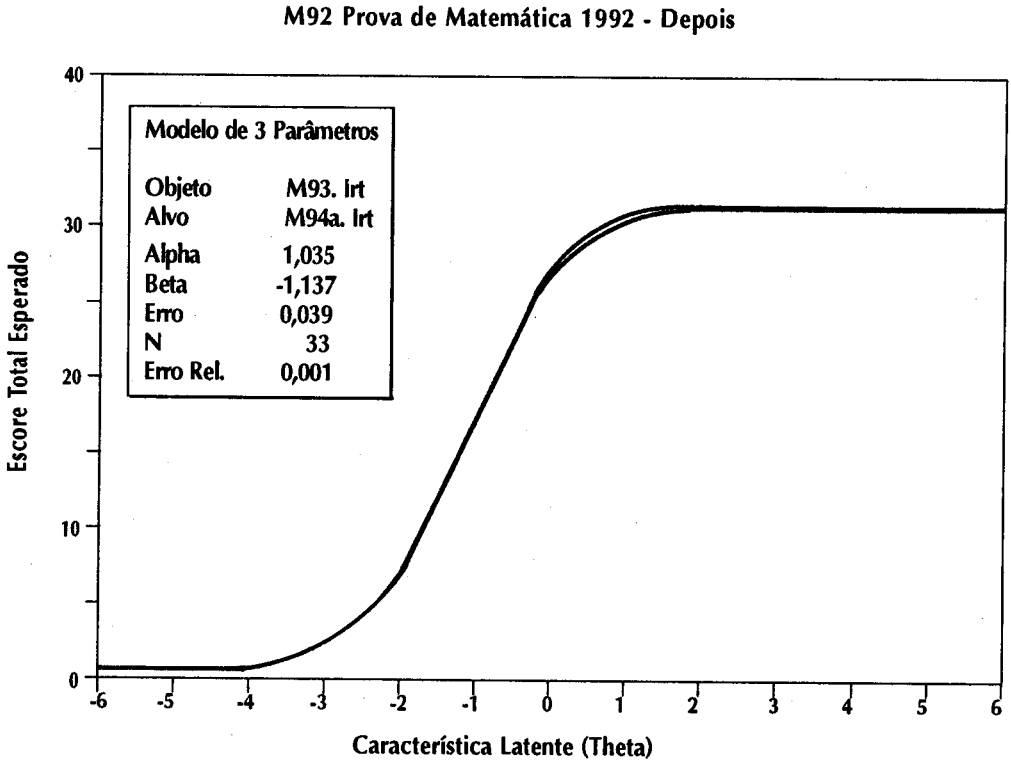

GRÁFICO 4B

dida θ é independente do número de itens. Portanto, pares de escores verdadeiros ζ_o e η_o correspondentes a determinado θ_o representam níveis de habilidade idênticos.

Na prática, a relação de equivalência para escores verdadeiros é estimada utilizando os parâmetros de item estimados para calcular as funções de característica das provas apresentadas acima. Pares de valores de ζ e η são computados para uma série de valores arbitrários de θ. Procuram-se valores para as constantes α e β, aplicadas a todos os parâmetros de item ***a*** e ***b***, que minimizam a média dos erros quadrados da diferença $\zeta - \eta$.

Escores Verdadeiros em 33 Itens Comuns a Duas Provas Após a Transformação Linear

Os coeficientes α e β do Gráfico 4B correspondem aos coeficientes da regressão linear apresentados anteriormente no Gráfico 3. Teoricamente, os novos coeficientes são mais consistentes porque levam em conta todos os parâmetros de item, ***a*** e ***b***. Na realidade, as duas soluções são muito parecidas. Com a aplicação da transformação linear definida por α e β, todos os parâmetros das duas provas se aproximam de uma identidade. Doravante, as duas provas devem ser consideradas equivalentes.

O uso da escala de habilidade θ dos modelos de resposta ao item (ou sua transformação) apresenta certos problemas de ordem prática. Embora pareça que os modelos da resposta ao item resolvam um grande número de problemas associados à tentativa de estabelecer a equivalência das provas, na prática trabalhamos com *estimativas* dos escores verdadeiros, e não com os escores verdadeiros, propriamente ditos, cujos valores verdadeiros permanecem desconhecidos. A tentativa de estabelecer equivalência sofre os efeitos do piso e do teto das provas. Escores abaixo de $\sum^{n}_{j=1} c_j$ não podem ser estimados. Esco-

res verdadeiros, correspondentes aos escores observados máximos nas duas provas, convergem automaticamente. Portanto, a conversão pode não cobrir toda a amplitude possível dos escores. No entanto, ao selecionar itens que satisfaçam as premissas do modelo da resposta ao item e ao calibrar esses itens com amostras de porte satisfatório, acredita-se que os valores das estimativas sejam próximos dos escores verdadeiros o suficiente para ganhar os importantes benefícios da equivalência.

Outro problema é que as estimativas de habilidade de examinandos de habilidade alta e baixa são sujeitas a doses relativamente grandes de erros de medida. Para o modelo logístico de três parâmetros, a variância dos erros de medida para examinandos de habilidade extrema pode ser facilmente 10 ou 100 vezes aquela dos examinandos mais típicos. Na prática, uma interpretação adequada das estimativas de habilidades para indivíduos, θ_i, requer que se leve em conta o erro associado à estimativa de θ. A grande variabilidade nos erros de medida das estimativas de θ pode criar problemas ao interpretar estatísticas tais como os coeficientes de regressão ou de correlação. Normalmente, as observações são ponderadas para levar em conta a maior variância dos erros nos extremos da prova.

ABSTRACT

Large scale evaluations often require alternative forms of the same test or tests of variable difficulty. In the first case, the forms are designed at the same level of difficulty but vary for unpredictable reasons. In the second, different forms are designed to vary in content and difficulty, often by a considerable degree. In these circumstances, use of a method to establish the comparatibility of results on multiple test forms becomes essential. Recent advances in psychometrics facilitate the task of placing all results on the same scale, taking advantage of item parameter invariance properties obtained through the use of item response models.

Referências Bibliográficas

ANGOFF, W. H. Scales, norms, and equivalent scores. In THORNDIKE, R. L. (Ed.), *Educational Measurement* (2. ed.). Washington, D. C.: American Council on Education, 1971.

GULLIKSEN, H. Theory of mental tests. New York: Wiley, 1950.

LORD, F. M. *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum, 1980.

LORD, F. M. Practical applications of item characteristic curve theory. *Journal of Educational Measurement,* 14, p. 117-138, 1977.

LORD, F. M., NOVICK, M. R. *Statistical theories of mental test scores.* Reading , MA: Addison-Wesley, 1968.

STOCKING, M. L., LORD, F. M. Developing a common metric in item response theory. *Applied Psychological Measurement,* 7, p.201-210, 1983.