Ministério do Planejamento, Orçamento e Gestão
Instituto Brasileiro de Geografia e Estatística – IBGE
Escola Nacional de Ciências Estatísticas

Textos para discussão
Escola Nacional de Ciências Estatísticas
Número 25

Da Teoria Clássica dos Testes para os Modelos de Resposta ao Item

Philip Ralph Fletcher (Professor Visitante)[1]

Rio de Janeiro

2010

[1] Pesquisador da WESTAT.

Instituto Brasileiro de Geografia e Estatística - IBGE

Av. Franklin Roosevelt, 166 - Centro - 20021-120 - Rio de Janeiro, RJ - Brasil

Textos para discussão. Escola Nacional de Ciências Estatísticas, ISSN 1677-7093

Divulga estudos e outros trabalhos técnicos desenvolvidos pelo IBGE ou em conjunto com outras instituições, bem como resultantes de consultorias técnicas e traduções consideradas relevantes para disseminação pelo Instituto. A série está subdividida por unidade organizacional e os textos são de responsabilidade de cada área específica.

ISBN 975-85-240-4126-6

Impressão

Gráfica Digital/Centro de Documentação e Disseminação de Informações – CDDI/IBGE, em 2010.

Capa

Gerência de Criação/CDDI

Fletcher, Philip Ralph

Da teoria clássica dos testes para os modelos de resposta ao item / Philip Ralph Fletcher. - Rio de Janeiro : Escola Nacional de Ciências Estatísticas, 2010.

37 p. - (Textos para discussão. Escola Nacional de Ciências Estatísticas, ISSN 1677-7093 ; n. 25)

Inclui bibliografia.
ISBN 975-85-240-4126-6

1. Teoria da resposta ao item. 2. Psicometria. I. Escola Nacional de Ciências Estatísticas (Brasil). II. Título. III. Série.

Gerência de Biblioteca e Acervos Especiais CDU 519.2:159.9.01
RJ/2010-10 EST

SUMÁRIO

RESUMO

A necessidade de uma teoria de testes e medidas é uma conseqüência do reconhecimento de que provas são por natureza falhas, porém, passíveis de aperfeiçoamento. Nas ciências do homem, ao contrário das ciências físicas, encontram-se peculiaridades que desafiam qualquer prática de medidas. Freqüentemente, conceitos fundamentados em acontecimentos sociais bem estabelecidos têm parco significado teórico, ao mesmo tempo que os conceitos de maior interesse teórico carecem de qualquer referência empírica clara. O que é corriqueiro nas ciências do homem acaba se tornando a regra em psicologia, onde as habilidades ou características que se pretendem elucidar geralmente não se manifestam diretamente, permanecendo subjacentes ou latentes. Elas se revelam apenas indiretamente, através da execução de certas tarefas, potencialmente relacionadas. Como não se pode medir ou representar diretamente a habilidade de uma pessoa em matemática, por exemplo, tenta-se aferir essa capacidade com uma prova, através do desempenho observado na execução de vários exercícios matemáticos ou "itens" selecionados especificamente para este fim. Em várias instâncias, pode-se até afirmar que um conceito psicológico alcança um grau de precisão em forma e definição apenas quando representado por uma medida específica. O texto apresenta um breve histórico sobre testes e formaliza o referencial teórico para a utilização da Teoria da Resposta ao Item – TRI.

Palavras-chave: TRI, Teoria da Resposta ao Item, variáveis latentes, variáveis não observáveis, escalas de proficiência

ABSTRACT

The need for a theory encompassing tests and measure scales is a natural consequence of the fact that tests are imperfect by nature. They can be improved, though. In humanities, as opposed to physical sciences, there are peculiarities that defy any attempt to define a measurement scale. Frequently, concepts based on well established social events have little theoretical significance, while, constructs of major theoretical interest have no empirical reference. What is frequent in humanities is basically the rule in Psychology, where abilities or characteristics one would like to study do not present themselves directly, lying subjacent or latent. They manifest themselves indirectly through the execution of certain tasks. As one cannot measure directly a person's ability in math, for example, one try to access this capacity with a test through the performance of several math exercises or item selected specifically for this purpose. Often, one can say that a psychology construct attains a certain precision level in form and characterization only when linked to a specific measure. This text presents a brief history of tests and formalizes the theoretical framework for the use of Item Response Theory – IRT.

Key-words: IRT, Item Response Theory, Latent variables, nin-observable variables, proficiency scales

1. Introdução

Qualquer pessoa que entra em contato com o sistema escolar se expõe a testes e medidas que pretendem aferir sua habilidade ou grau de domínio em determinada área curricular. É comum, nessas ocasiões, o indivíduo logo se surpreender pelos resultados obtidos. Vez por outra, sente-se vindicado por uma experiência bem sucedida, assim como, em outra ocasião, também pode se sentir traído, quando leva bomba. Num período de tempo relativamente curto, um único indivíduo pode sofrer experiências inusitadas, de maior ou menor grau de sucesso relativo. De fato, é isto que registram as notas de professor ao longo do semestre: normalmente, a ordem dos alunos estabelecida pelos resultados de uma prova numa primeira instância não coincide com a ordem verificada após a administração de uma segunda prova. Essa variabilidade de resultados chama a atenção para um problema bastante sério. Ao tentar medir a habilidade de um examinado, sob condições onde sua habilidade permanece inalterada, dois instrumentos de medidas perfeitas deveriam fornecer o mesmo resultado para o indivíduo examinado. Sendo as diferenças nos resultados das duas aplicações de prova incompatíveis com o propósito imediato de aferir e representar uma habilidade considerada estável, deve-se concluir que existem erros ou inconsistências nos procedimentos de prova.

Ao aplicar, no mesmo grupo de examinados, duas provas que medem a mesma habilidade, obtêm-se duas distribuições de resultados diferentes. A discrepância entre as duas distribuições depende, entre outros fatores, da diferença no grau de dificuldade relativa dos dois testes. Qual das duas distribuições deveria ser escolhida para representar a distribuição de habilidades? Infelizmente não existe nenhuma justificativa para escolher uma ou outra dessas duas distribuições. Este exemplo mostra que as distribuições dos resultados de prova não oferecem qualquer informação sobre a distribuição das habilidades. O problema decorre da desigualdade das unidades de medida das duas provas, que variam de acordo com a dificuldade dos itens. Muito pelo contrário, uma verdadeira escala de habilidade seria aquela que proporcionasse medidas idênticas em provas não equivalentes as que avaliam a mesma habilidade. Neste caso, sim, mesmo que os resultados de prova venham de instrumentos diferentes, a expectativa da estimativa de habilidade do examinado deveria ser a mesma em cada um dos dois casos.

A necessidade de uma teoria de testes e medidas corre como conseqüência do reconhecimento de que as provas são por natureza falhas, porém, passíveis de aperfeiçoamento. Nas ciências do homem, ao contrário das ciências físicas, encontram-se peculiaridades que desafiam qualquer prática de medidas. Freqüentemente, conceitos fundamentados em acontecimentos sociais bem estabelecidos têm parco significado teórico, ao mesmo tempo que os conceitos de maior interesse teórico carecem de qualquer referência empírica clara. O que é corriqueiro nas ciências do homem acaba se tornando a regra em psicologia, onde as habilidades ou características que se pretendem elucidar

geralmente não se manifestam diretamente, permanecendo subjacentes ou latentes. Elas se revelam apenas indiretamente, através da execução de certas tarefas, potencialmente relacionadas. Como não se pode medir ou representar diretamente a habilidade de uma pessoa em matemática, por exemplo, tenta-se aferir essa capacidade com uma prova, através do desempenho observado na execução de vários exercícios matemáticos ou "itens" selecionados especificamente para este fim. Em várias instâncias, pode-se até afirmar que um conceito psicológico alcança um grau de precisão em forma e definição apenas quando representado por uma medida específica.

Quando se propõe a examinar características não observáveis, cujo potencial permanece apenas subjacente, a relevância e utilidade prática de uma teoria de medidas acabam se tornando patentes. Logo de saída, é impraticável aplicar a mesma medida várias vezes em sujeitos humanos. Em matéria de provas, pode-se aplicar o mesmo instrumento uma ou duas vezes, mas ao continuar insistindo em sucessivas aplicações, as respostas do examinado se alteram dramaticamente, simplesmente por causa da fatiga ou dos efeitos da prática. Outra diferença que desafia a prática dos testes advém da necessidade de examinar um grande número de pessoas, geralmente todas ao mesmo tempo, mas com o propósito de chegar a algumas conclusões pertinentes sobre indivíduos isolados e em relação ao grupo. Dificuldades lógicas e de metodologia estatística, ao formular inferências simultâneas sobre indivíduos e grupos, introduzem complexidades que dificilmente encontram eco nas ciências físicas, onde geralmente é possível observar apenas um objeto ou evento de cada vez.

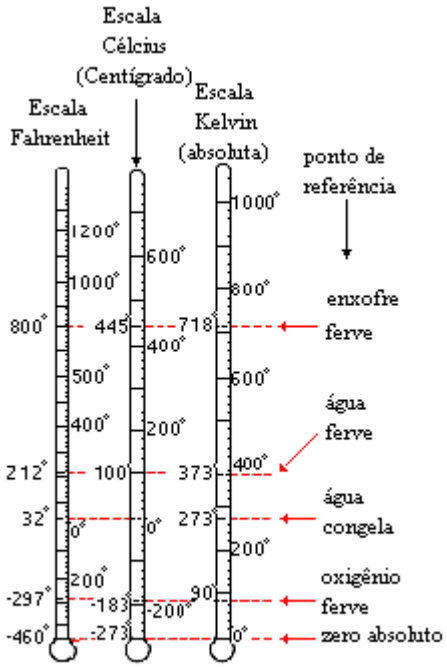
As medidas em psicologia, assim como nas outras ciências, iniciam-se por um processo de relacionar elementos do mundo real com os elementos ou formulações de um sistema de lógica abstrata ou modelo teórico. Através da definição semântica dos elementos básicos dessa teoria, procura-se levar adiante na prática das medidas uma explicação do conceito original. Freqüentemente, o desenvolvimento de um instrumento de medidas devidamente adequado para os propósitos previstos acaba sendo um ato tão importante pelo que ensina sobre o assunto em consideração quanto a atividade subseqüente de medir com a aplicação do instrumento. O êxito alcançado ao medir certa propriedade, por si só, demonstra um profundo conhecimento da característica em questão.

2. Referencial teórico

O conceito de temperatura oferece um exemplo prático que se elabora e enriquece ao levar adiante o processo de medidas. A temperatura se anuncia na forma de um conceito abstrato definido utilizando os termos mais simples e precisos possíveis. Uma definição teórica da temperatura afirmaria que ela é a energia cinética encontrada na moção aleatória das moléculas de uma substância. A definição conceitual sugere procedimentos a seguir para formar medidas que representam esse conceito. Utiliza-se um líquido que expande e contrai com a variação da energia cinética presente para criar um instrumento que mede a temperatura. A leitura do termômetro em contato com a substância se torna uma medida de sua temperatura. A rigor, não se observa diretamente a propriedade da temperatura mas apenas sua manifestação no termômetro. Embora seja o termômetro que é medido, costuma-se dizer que a temperatura da substância foi tomada ou registrada. Parte-se, então, para detalhar denotações específicas atribuídas ao termo. Com o aumento da energia cinética, um termômetro registra valores sucessivos mais altas ao longo de sua escala. Uma série de pontos de referência relacionam acontecimentos específicos com um único valor de escala, facilitando a interpretação de sua métrica em termos do potencial cinético (Figura 1).

Figura 1

Escalas Lineares, Invariantes e Conversíveis, com Pontos de Referência Bem Definidos



Escalas de Temperatura

Virtualmente todas as medidas abrigam algum grau de erro. Assim, o próximo passo no desenvolvimento de uma medida envolve uma consideração do grau de erro inerente ao instrumento e aos procedimentos de medida. Para examinar esta questão, pode-se usar dois termômetros: um demarcado em graus fahrenheit e o outro em graus célsius, por exemplo. Ao saber que os graus fahrenheit traduzem os graus célsius segundo uma função matemática exata (F° = 9/5 C° + 32°), é possível realizar experiências com os dois tipos de termômetro para avaliar o erro de medida implícito no ato da leitura de temperatura.

Prossegue-se a verificar que os procedimentos de medida dão resultados satisfatórios quando aplicados a diferentes substâncias sob diferentes condições. Procede-se a fazer comparações baseadas nas diferenças das medidas observadas. Para que isto seja válido, espera-se que a unidade de diferença tenha o mesmo significado em toda a extensão da escala de temperatura. A equivalência ou invariância das diferenças obtidas com os procedimentos de medida numa grande variedade de circunstâncias constitui o princípio que viabiliza a comparação da energia cinética contida em diferentes substâncias sob diversas condições específicas. O estudo da energia cinética se tornaria mais difícil caso os instrumentos e os procedimentos de medida interagissem de alguma forma com as substâncias medidas. Nossos objetivos seriam seriamente comprometidos caso os resultados dependessem do termômetro em particular que foi aplicado. Seria desconcertante, também, descobrir que os termômetros fornecem resultados adequados para substâncias como água, mas não para enxofre, por exemplo. Nessas circunstâncias, o princípio da invariância das diferenças teria sido violado, inviabilizando as comparações.

Embora existam algumas similaridades entre os modelos de medidas nas ciências físicas e os modelos usados nas ciências humanas, vale a pena notar mais algumas diferenças. Em primeiro lugar, a definição teórica de um conceito como a temperatura alcança um maior grau de aceitação que a maioria dos conceitos aplicados nas ciências humanas. As diferenças nas definições podem levar a uma maior variedade de medidas dos conceitos em sociologia ou psicologia do que nas ciências físicas. Em segundo lugar, a

escala usada para medir a temperatura é uma questão de consenso aceita pela grande maioria dos cientistas, ao mesmo tempo que praticamente nenhuma escala é universalmente aceita em sociologia ou psicologia. Finalmente, o grau de erro da medida é maior nas ciências do homem do que nas ciências físicas. Não se consegue medir conceitos em psicologia ou sociologia com alto grau de precisão aplicando as técnicas disponíveis hoje. Torna-se necessário o uso de modelos de medidas explícitos para entender melhor as relações entre as variáveis medidas e os conceitos subjacentes a elas (Bollen, 1989; capítulo 6).

As medidas servem para descrever objetos do mundo real com números. Identificamos os atributos a serem medidos, selecionamos os instrumentos e as unidades apropriados para sua representação, aplicamos os conceitos de medida ao descrever as observações e invocamos as medidas para comunicar nossas idéias para outras pessoas. A medição seria portanto um processo idealizado para designar números (escores, medidas ou resultados) que se tornam a representar propriedades específicas de unidades experimentais de tal forma que se caracterize e preserve relacionamentos específicos encontrados no mundo real (Lord e Norvick, 1968; 17). A designação dos valores numéricos segue uma lógica que permite sua análise por operações matemáticas, seguindo determinadas regras específicas. Uma álgebra ou cálculo, normalmente de uma ordem bastante sofisticada, facilita manipulações das construções elementares, permitindo deduções do modelo. O objetivo não seria de explicar qualquer evento, comportamento ou resposta isolado, porque qualquer um destes seria um reflexo de múltiplos determinantes conflitantes. Seria melhor afirmar que a intenção das medidas é de representar a consistência no comportamento ou numa série de observações. As consistências são resumidas geralmente na forma de escores totais ou de sub-escores que sim, freqüentemente, refletem um único determinante específico. Dessa maneira, passamos do nível de comportamentos discretos ou de observações isoladas para o nível próprio das medidas.

A formulação matemática do processo de medida representa uma enorme economia de expressão e ainda uma melhoria apreciável em termos de precisão relativa a seu equivalente discursivo, expresso em linguagem verbal. No entanto, essa formulação, ao abstrair-se da realidade comportamental e ao tornar-se cada vez mais rigorosa e explícita, tende a perder a força de seu significado conotativo, empalidecendo o conteúdo conceitual. Por essa razão, o real significado da medida precisa ser constantemente ressuscitado e perseguido. A preocupação com o significado, utilidade e legitimidade do uso dos resultados é necessária não apenas para sustentar a interpretação de uma medida mas ainda para justificar seu uso em situações aplicadas. Refere-se assim ao esforço para estabelecer a validade dos testes e medidas. A validação da medida requer não apenas uma explicação de seu significado ostensivo, mas também uma consideração dos juízos de valor nela embutidos e ainda sua relevância prática em determinados contextos operacionais, sobretudo uma apreciação das conseqüências sociais de seu uso em processos de tomada de decisão.

A questão fundamental, que fatalmente surge no caminho da validação, é se o significado da medida é relevante apenas num determinado contexto específico ou se, de alguma forma ou outra, ela pode ser generalizada para outros contextos. Embora existam diversas maneiras de acumular evidências para sustentar inferências neste sentido, seus meios são essencialmente os procedimentos da pesquisa científica. Procura-se estabelecer se a medida demonstra as mesmas propriedades invariantes e configurações de relacionamento observadas em diferentes grupos populacionais, sob uma grande diversidade de condições objetivas. As possibilidades lógicas não são assim ilimitadas (Messick, 1993):

- Como é que a pessoa se desempenha num *universo* de situações? Examina-se o conteúdo da prova em relação ao conteúdo do domínio de referência.

- Até que ponto a prova mede uma consistência observada? Verifica-se como as pessoas respondem aos itens ou tarefas, examinando as relações entre respostas para evidência de sua consistência *interna*.
- Como é que a pessoa se posiciona em relação a alguma outra variável diferente da prova? Examina-se a relação entre os resultados de prova e outras medidas de características ou de experiência para evidência de estrutura *externa*.
- Quão generalizáveis, estáveis ou sensíveis são os resultados da prova? Procuram-se diferenças na estrutura dos resultados ao longo do tempo, entre grupos e situações e como resposta a intervenções experimentais que manipulam o currículo, os procedimentos didáticos ou terapêuticos.
- Será que o uso da prova provoca conseqüências sociais ou levanta questões tratadas em políticas públicas? Examinam-se as conseqüências sociais da interpretação e uso dos resultados de prova para determinados fins, considerando não apenas seus efeitos previsíveis mas também seus efeitos colaterais não intencionados, especialmente seu viés, abuso, impacto adverso ou os efeitos sociais mais sutis associados ao seu uso habitual e recorrente.

Todas essas formas de evidência pesam na interpretação e uso das medidas de prova. A variedade das formas de validade não pretende expressar uma divergência de opções alternativas, mas, sim a convergência das facetas complementares de um mesmo conceito, consistentes entre si em termos de seu embasamento lógico e operacional. Torna-se necessária a procura de evidências de generalidade para inferir quão extensivo ou circunscrito será o significado dos resultados de prova. Embora seja impossível provar a validade de um instrumento, usando os métodos da ciência, é possível desenvolver forte apoio em seu favor.

As teorias de testes e medidas surgem no meio dessa procura de um quadro de referência que permita relacionar simultaneamente formulações matemáticas cada vez mais rigorosas com as propostas psicológicas mais abrangentes. Veja-se, portanto, que a ciência avança tanto pela teoria quanto pelos meios de observação. Nesse caminho, encontramos interpretações instrumentalistas e realistas lado a lado. As construções mais abstratas são recursos heurísticos que visam organizar relações observadas, sem necessariamente pressupor que existam entidades reais subjacentes a elas. Do outro lado, os dados usados como evidência são manifestações de características comportamentais e não as meras manifestações de construtos teóricos. Na tentativa de armazenar fato, significado e valor para sustentar a interpretação e uso dos testes e medidas, o percurso da validação é invariavelmente o mesmo.

A relação entre os âmbitos teórico e comportamental se estabelece através de variáveis observáveis escolhidas com o propósito específico de se servirem como medidas das construções teóricas. No caso específico das provas, as variáveis se referem ao desempenho demonstrado pelo examinado numa série de tarefas relacionadas com o atributo postulado como conceito teórico. A variável observável opera como medida do conceito teórico se seu valor numérico aumenta monotonicamente junto com a presença do postulado teórico. Desta forma, o êxito alcançado com a execução de uma seqüência de tarefas específicas é considerado uma manifestação da presença do atributo investigado. Numa prova de múltipla escolha, o desempenho medido diretamente consiste das respostas do examinado aos itens de prova, o que demonstra sua habilidade de reconhecer e identificar o procedimento o conceito solicitado. A resposta pode ser representado pelo valor de 1 para acertos e 0 para erros. A partir de cada item de prova, acumula-se um conjunto de respostas na forma de uma seqüência de uns e zeros, 001101011010 ... , o que corresponde ao vetor de respostas x = (x1, x2, ... , xL), onde L é o número total de itens na prova.

Propõem-se diversas maneiras para resumir as informações contidas nesta seqüência de algarismos por uma única medida sintetizando o desempenho do examinado. Uma escala de escores é uma seqüência de valores numéricos ordenados, proposta para refletir progressíveis níveis de desempenho ou de habilidade. A maioria dos testes em psicologia especificam um modelo de medidas que incorpora uma escala 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 e a sua ordem relativa, mas também às diferenças relativas entre valores encontrados ao longo da escala. O intervalo é, portanto, uma distância única que relaciona quaisquer dois pontos encontrados ao longo da escala. A diferença de unidade de 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.

O número total acertos na prova pode ser estipulado como a regra numérica a ser aplicada no ato da mensuração. Nos termos mais gerais, a escala de uma prova pode ser definida por uma fórmula de escores X, onde

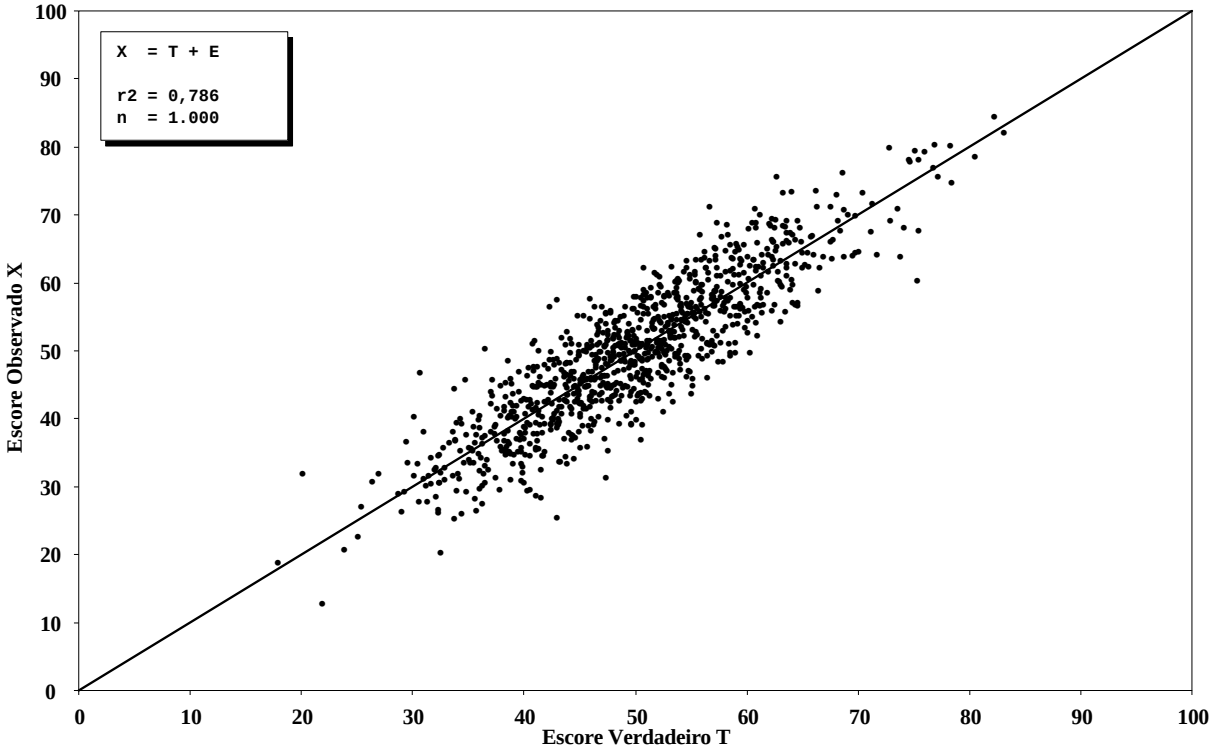
$$X_i = \sum_{j=1}^{L} w_j x_{ij} \qquad \textbf{(1)}$$

sendo a medida xij = 1 quando o item j for acertado pela pessoa i e xij = 0 quando o contrário. Surge como dúvida qual seria o valor da ponderação wj a ser atribuída a cada item j = 1, 2, ... , L. Quando for sempre w = 1, a fórmula proporciona o total de acertos, enquanto w = 1 / L apresenta o resultado na forma da proporção de acertos. Onde o valor de w permanece constante em todos os itens, obtém-se o modelo linear da chamada teoria dos valores verdadeiros ou, mais simples, a teoria clássica de testes.

A teoria clássica tem suas origens no modelo de escore verdadeiro e de erro apresentado pelo psicólogo britânico Charles Spearman (1863-1945) em 1904. Fascinado pelo conceito de correlação, Spearman publicou uma série de ensaios matemáticos onde argumentou que os resultados das provas são imperfeitos por natureza e, portanto, a correlação observada entre duas medidas falhas seria fatalmente inferior à correlação realmente existente, caso fossem conhecidos seus "valores objetivos verdadeiros". Ao explicar o significado dos termos medidas falhas e valores objetivos verdadeiros, Spearman criou os fundamentos da teoria clássica. Seu alcance foi enorme, predominando a disciplina dos testes e medidas ao longo dos próximos 50 anos. Provavelmente, ainda hoje, a maioria dos técnicos da área de testes e medidas continua aplicando a teoria clássica em alguma forma ou outra, embora nem sempre tenha plena consciência disso. Guilford (1936) e Gulliksen (1950) fazem apresentações da teoria clássica, quase sempre citadas. O leitor interessado na literatura a seu respeito ainda pode consultar as referências citadas por Stanley (1971) e as exposições apresentadas em Lord & Novick (1968), Cliff (1979), de Gruijter & van der Kamp (1984) e Crocker & Algina (1986). No Brasil, a teoria clássica foi introduzida por Vianna (1973).

Gráfico 1

Modelo Clássico: Escore Verdadeiro Distorcido pelo Erro Aleatório e Independente



O modelo clássico parte do pressuposto de que o escore de uma prova é por natureza falha e contém alguma parcela de erro. Logo, se uma parte do resultado é de erro, o restante deve ter sua base real ou "verdadeiro". Dessa forma, o escore observado *X*, numa prova qualquer, representa a soma de um componente de escore verdadeiro *T* (true score) e um componente de erro *E*, todos apresentados na mesma escala de unidades que os escores observados (Gráfico 1). Generalizando, para a pessoa *p* e a forma de teste *f*:

$$X_{pf} = T_p + E_{pf}. \qquad (2)$$

Ao especificar uma fórmula de escore baseada na soma dos valores *1* atribuídos aos acertos e 0 aos desacertos, *Tp* será delimitado por *0* e *L*, onde *L* é o escore máximo possível na prova. O escore verdadeiro da pessoa é um constante em particular que não varia de acordo com a forma da prova aplicada, enquanto o erro ocorre simplesmente ao acaso. Em tese, o erro resulta da impossibilidade de incluir um número infinito de itens na prova, aplicar a provas no indivíduo um número infinito de vezes ou submeter as respostas a um número infinito de examinadores. Sob a perspectiva clássica, o erro não se relaciona com o escore verdadeiro e não se relaciona com o erro que ocorre em qualquer outra medida da mesma característica. O erro é tido como *aleatório, sem viés* e *independente dos escores verdadeiros*.

O escore verdadeiro e o erro, evidentemente, não são observáveis. Observam-se apenas amostras de comportamento na execução de uma série de tarefas, e a partir dessas observações se faz inferências sobre os conceitos de escore verdadeiro e de erro. As inferências que podem ser feitas dependem em parte da natureza das amostras de comportamento observadas. O ponto central é que o universo da experiência se caracteriza por uma dimensão de conteúdo, além das tarefas ou tentativas realizadas e dos juízes ou bancas que classificam os resultados entre acertos e equívocos. O desempenho verificado reflete em parte a variação do conteúdo específico selecionado para representar o universo.

O conteúdo desse universo pode receber uma definição estreita, quando as palavras de uma prova de vocabulário vêm de uma única área de conhecimento, por exemplo, biologia, ou ampla, quando representam diversas áreas. O universo pode ser definido apenas em termos de conteúdo, ou pode ser também limitado em termos de formato, por exemplo, quando se exige que o examinado identifique como sinônimos ou antônimos pares de palavras apresentados fora de contexto. Assim que a definição do universo se amplia, qualquer tarefa única ou amostra restrita de tarefas se torna menos apta para representar o universo. Portanto, os resultados de uma prova se referem ao escore verdadeiro de um universo particular, e dependem tanto da definição do universo quanto da amostra de itens incluída no instrumento.

Partindo de uma amostra de itens considerada representativa do universo escolhido para investigação, define-se o escore verdadeiro como o valor médio alcançado por medidas idênticas quando o número de medidas aumenta sem limite. A noção de "medidas idênticas" implica que seja possível medir o desempenho de uma pessoa reiteradamente sem alterá-la, uma condição sem correspondência no mundo real. No entanto, conceitualmente, é útil pensar assim, com a pessoa permanecendo inalterada ao longo da observação replicada. Sob essas condições ideais, formas de prova paralelas *f, g, h*, ... , medidas alternativas construídas das mesmas especificações, produzem distribuições de escores observados idênticas em amostras muito grandes, covariam igualmente entre si e covariam igualmente com qualquer outra medida *Z* que não seja uma das formas paralelas. Em símbolos (Stanley, 1971),

$$\begin{aligned} F(X_f) &= F(X_g) = F(X_h) = \dots \\ \sigma_{x_f x_g} &= \sigma_{x_f x_g} = \sigma_{x_f x_g} = \dots \\ \sigma_{x_f z} &= \sigma_{x_f z} = \sigma_{x_f z} = \dots \end{aligned} \tag{3}$$

Se fosse possível medir várias vezes a mesma pessoa *p*, embora utilizando uma forma paralela diferente em cada instância, a média dos erros (*Ep1 , Ep2 , ...*) se aproximariam do limite de 0 com o aumento infinito das medições. A expectativa ou valor médio *Ef* tende a zero com a administração de um número infinito de formas paralelas *f*:

$$E_f E_{pf} = 0. \tag{4}$$

Isto segue como conseqüência dos erros definidos como desvios aleatórios em torno dos escores verdadeiros, sendo igualmente provável a ocorrência tanto de erros positivos como de negativos. No limite, a média dos escores observados da pessoa é igual a seu escore verdadeiro, isto é: $E_f X_{pf} = \overline{T_p}.$ Em termos operacionais, o escore verdadeiro é o escore médio obtido pela pessoa a partir da aplicação de um número infinito de formas paralelas.

A mesma linha de raciocínio que caracteriza as medidas repetidas de um indivíduo é aplicada à média de um grupo. Para uma população infinita de pessoas, examinada com qualquer forma do teste em particular, a média dos erros obtidos (*E1f , E2f* , ...) se aproxima de *0*. Em símbolos,

$$E_p E_{pf} = 0, \tag{5}$$

onde *Ep* representa o valor esperado para pessoas. No limite, a média dos escores observados de pessoa é igual ao escore verdadeiro do grupo, $E_p X_{pf} = \overline{T_p}.$ Portanto, o escore verdadeiro do grupo se define pelo escore médio de seus membros. Assim que o número de observações aumenta, a média observada se aproxima da média dos escores verdadeiros e se torna uma estimativa não enviesada da média dos escores verdadeiros.

Como conseqüência dos pressupostos iniciais apresentados acima, os erros provenientes da aplicação da forma *f* são independentes dos escores verdadeiros subjacentes a *Xf* e independentes do erro encontrado em qualquer outra forma da prova. A

independência de *Ef e T*, no teste X, antecipam a relação de variância fundamental à teoria clássica:

$$\sigma^2_{x_f} = \sigma^2_T + \sigma^2_{T_f}. \quad (6)$$

Na ausência de uma correlação entre os valores verdadeiros e de erro, o termo da covariância é ignorado e desaparece dessa relação. A variância total seria, portanto, a soma das variâncias dos valores verdadeiros e de erro.

Embora os aplicadores de provas conheçam apenas os escores observados, parece óbvio que o que realmente gostariam de saber são os escores verdadeiros. Assim, uma pergunta importante é qual seria o grau de proximidade entre os escores observados e os verdadeiros? O coeficiente de correlação, que expressa o grau de relação entre escores observados e verdadeiros, seria um índice dessa fidedignidade:

$$\rho_{XT} = \frac{\sigma^2_T}{\sigma_X \sigma_T} = \frac{\sigma_T}{\sigma_X}. \quad (7)$$

O quadrado desse valor é o coeficiente de determinação da regressão de *T* em *X*, igual à parcela da variância dos escores observados compartilhada pelos escores verdadeiros (veja o coeficiente de determinação r2 = 0,786 apresentado no Gráfico 1). No entanto, esta formulação é de interesse apenas teórica porque não se pode observar diretamente os escores verdadeiros, assim como também não se pode obter todos os possíveis escores observados para cada pessoa examinada. No entanto, quando dois instrumentos satisfazem os requisitos de formas de prova paralelas, é possível estabelecer uma identidade matemática que relaciona a correlação entre escores observados e escores verdadeiros ρ_{XT} e a correlação de escores observados em duas formas paralelas $\rho_{X_f X_g}$. Segundo a teoria clássica, as formas seriam paralelas quando cada pessoa tem o mesmo escore verdadeiro nas duas formas da prova e quando as variâncias das duas formas são iguais, conforme anunciado em (3). As formas teriam, como conseqüência, escores observados com médias e variâncias iguais, a partir das quais pode-se inferir que as variâncias dos erros também seriam iguais. Pressupõe-se, ainda, que as formas de prova coincidem em termos de formato e conteúdo substantivo.

Nestas circunstâncias, examina-se a correlação entre duas formas paralelas *f* e *g*:

$$\rho_{X_f X_g} = \frac{\sigma_{X_f X_g}}{\sigma_{X_f} \sigma_{X_g}} = \frac{\mathrm{E} X_f X_g}{\sigma_{X_f} \sigma_{X_g}} = \frac{\frac{1}{N}\sum X_f X_g}{\sigma_{X_f} \sigma_{X_g}}. \quad (8)$$

Ao perceber que *Xf* e *Xg* representam a soma dos escores verdadeiros e de erro e que, no denominador, os dois desvios padrões são iguais, essa correlação pode ser escrita:

$$\begin{aligned} \rho_{X_f X_g} &= \frac{\mathrm{E}(T + E_{X_f})(T + E_{X_g})}{\sigma_{X_f X_g}} \\ &= \frac{\mathrm{E}T^2 + \mathrm{E}TE_{X_f} + \mathrm{E}TE_{X_g} + \mathrm{E}E_{X_f} E_{X_g}}{\sigma_{X_f X_g}} \\ &= \frac{\frac{1}{N}(\sum T^2 + \sum TE_{X_f} + \sum TE_{X_g} + \sum E_{X_f} E_{X_g})}{\sigma^2_{X_f}} \\ &= \frac{\sigma^2_T + \sigma_{TE_{X_f}} + \sigma_{TE_{Xg}} + \sigma_{E_{X_f} E_{Xg}}}{\sigma^2_{X_f}}. \end{aligned} \quad (9)$$

Sendo que os erros não se correlacionam com os escores verdadeiros e não se correlaciona entre se, os três termos de covariância no numerador se tornam nulos, deixando:

$$\rho_{X_f X_g} = \frac{\sigma_T^2}{\sigma_X^2} = \rho_{XT}^2 . \qquad (10)$$

O quadrado da correlação entre escores observados e escores verdadeiros é, portanto, igual à correlação entre medidas paralelas. Este valor é conhecido como o coeficiente de fidedignidade do teste, um dos principais índices do modelo clássico. Veja-se que o coeficiente de determinação de r2 = 0,786, apresentado no Gráfico 1, corresponde ao coeficiente de fidedignidade $\rho_{XT}^2 = 0{,}786$.

Ao interpretar as informações sobre fidedignidade, é importante notar que o coeficiente definido para determinado conjunto de medidas de teste tem relevância apenas teórica. Ele representa o valor que seria alcançado na aplicação de formas realmente paralelas. Em qualquer situação aplicada, os escores de formas reais substituem as medidas consideradas formalmente paralelas. Na realidade, aproximam-se medidas paralelas com a mesma forma de prova aplicada em duas ocasiões (o que proporciona um coeficiente de estabilidade), duas formas alternativas do mesmo teste, administrados simultaneamente (coeficiente de equivalência), ou ainda que relaciona a variância do escore da prova em seu conjunto com a covariância dos itens que o compõem (coeficiente de consistência interna). Na medida que estes procedimentos apenas aproximam o ideal das medidas paralelas, o coeficiente de fidedignidade calculado a partir de sua base seria apenas uma aproximação da proporção de variância comum a escores observados e verdadeiros.

O coeficiente de fidedignidade serve diversas finalidades no âmbito da teoria clássica. O coeficiente é um índice da eficácia de um instrumento de medidas, revelando até onde os resultados obtidos com determinado método de mensuração são replicáveis. A partir da relação $\sigma_T^2 = \sigma_X^2 (1 - r_{X_f X_g})$, o coeficiente de fidedignidade é usado para estimar o erro padrão dos escores e para estabelecer zonas de confiança em torno dos valores observados. Na presença dos erros de medida, as correlações entre escores verdadeiros seriam sempre mais altas que as observadas entre medidas falíveis. O coeficiente de fidedignidade serve para estimar o efeito dos erros de medida sobre as correlações e corrigir sua atenuação, o problema que inspirou a obra de Spearman no início do século. A teoria clássica demonstra que a variância dos escores verdadeiros aumenta junto com o quadrado do número de itens no teste, ao mesmo tempo que a variância dos erros aumenta apenas linearmente com o número de itens. Essa relação explica o aumento progressivo da fidedignidade do teste junto com o aumento do número de itens.

A maneira mais prática de produzir testes mais fidedignos é simplesmente de aumentar o número de itens. A fórmula de profecia de Spearman-Brown utiliza o coeficiente de fidedignidade para estimar o número de itens que seria necessário para alcançar determinado nível de fidedignidade. Tipicamente, uma prova curta, de apenas 15 itens, poderia ter um coeficiente de fidedignidade de 0,50; uma prova de 30 itens bem elaborados, uma fidedignidade de 0,80; e 85 itens, num teste aprimorado ao longo de vários anos, um coeficiente de 0,90. Finalmente, deve-se mencionar que existem *recomendações* que definem os níveis de fidedignidade a serem alcançados em determinadas situações práticas. Assim, em investigações preliminares, um coeficiente de fidedignidade entre 0,50 e 0,60 seria considerado adequado para análises exploratórias; em pesquisas com objetivos principalmente teóricos, não existiriam motivos para ultrapassar um patamar em torno de 0,80; no entanto, nas aplicações que podem afetar as oportunidades de vida dos examinados, 0,90 deve ser considerado o mínimo aceitável e 0,95 o valor desejado (Nunnally, 1967).

É importante reconhecer que as principais equações do modelo clássico são simples identidades tautológicas que seguem da definição de *T* e *E* anunciada em (2). São equações que não podem ser falsificadas por qualquer conjunto de dados. Deve ser notado também que σ_T^2, σ_E^2 e σ_{XT}^2 não podem ser estimados sem antes pressupor que os erros de medida encontrados nas formas de prova *Xf* e *Xg* não se correlacionam entre si e não se correlacionam com os escores verdadeiros. Nessas circunstâncias, a teoria clássica ainda conceitualiza os escores observados, verdadeiros e de erro como valores contínuos, desinibidos por quaisquer limites naturais, embora na prática fique patente a natureza discreta das distribuições de escores observados. Persiste ainda com a ficção de que o grau de dificuldade do teste não interfere com o grau de erro, e que o número finito de itens não restringe a variância dos escores nos limites superior e inferior do teste.

O coeficiente de fidedignidade, único para toda a distribuição, pressupõe uma homogeneidade na variância dos erros em qualquer nível de habilidade, embora isto não pareça ser plausível na prática. Aceita-se a veracidade dessa representação, sem dispor de procedimentos de avaliação que permitiriam demonstrar sua relevância. No mundo real, é pouco provável que a medida de eficácia preferida pela teoria clássica seja útil como índice de precisão nos extremos da distribuição de escores observados. Em todo caso, o coeficiente depende das variâncias dos escores verdadeiros e de erro encontrados numa população específica qualquer. Portanto, a fidedignidade definida se refere apenas a uma população concreta e particular. Isto leva Samejima (1977) a afirmar categoricamente que a "fidedignidade é um conceito morto na teoria dos testes, tendo em vista que varia de um grupo de pessoas para outro e as possibilidades para sua generalização são estreitamente limitadas" (p. 243).

Finalmente, deve-se notar que a teoria clássica preconiza a montagem de um número ilimitado de formas de prova rigorosamente paralelas. Por definição, cada pessoa tem o mesmo valor verdadeiro e a mesma variância de erro em cada forma paralela. Apesar do conceito ocupar uma posição de destaque na teoria clássica, a própria teoria nada ou pouco informa sobre a montagem dessas formas paralelas. Na prática, as formas de prova não existiriam em número ilimitado. Mesmo que existissem, não poderiam ser administradas sem antes provocar alterações permanentes no examinado. Nestas circunstâncias, as noções de escore verdadeiro e de variância de erro não passam de meras noções conceituais. Eventualmente, poder-se-ia imaginar que uma ou outra das formas existentes representem amostras tiradas de uma infinidade de formas de prova, mas obviamente essa infinidade é apenas outra ficção conveniente, que não encontra qualquer correspondência na vida real.

Num artigo influente, Osburn (1968) desmistifica as fundações da noção de amostragem subjacente à teoria clássica ao afirmar:

> *Poucos especialistas em medidas questionariam a premissa de que o objetivo fundamental do uso de testes de desempenho cognitivo seja sua generalização. No entanto, de fato, os procedimentos atuais usados na construção dos testes de desempenho não fornecem uma base sólida que permitiria a generalização para um universo de conteúdo bem definido. Na pior das hipóteses, os testes de desempenho representam coleções arbitrárias de itens, jogados juntos quase ao acaso. Na melhor, os testes contêm itens considerados relevantes e representativos em algum universo de conteúdo mal definido, elaborado por especialistas de currículo. Em nenhuma dessas hipóteses pode-se afirmar que exista uma base para generalizações. Isto ocorre porque o método usado na geração dos itens e os critérios de escolha dos itens não podem ser formulados em termos operacionais. [Osburn, 1968: 95].*

A mesma conclusão é compartilhada por Loevenger (1965) ao notar que:

> *o conceito de população (universo) implica que, em princípio, pode-se cadastrar, revelar ou enumerar todos seus membros possíveis, mesmo que a população (universo) seja infinita e o cadastro nunca será concluído. No entanto, é inconcebível que qualquer sistema pudesse enumerar todos os testes (itens) possíveis. Não existe um princípio para sua geração. [Loevinger, 1965: 147; grifo nosso].*

De fato, ninguém obteve um item por amostragem. Todos os itens são construtos.

Na realidade, ninguém sabe se o modelo clássico se aplica a qualquer teste. Muito do que se afirma na guisa da teoria clássica passa sem possibilidades de verificação, na forma de simples identidades tautológicas. Aceita-se sua veracidade sem dispor de procedimentos de avaliação que permitiriam demonstrar sua suposta relevância no mundo real. Pressupõem-se verdadeiras algumas de suas pressuposições fundamentais, mesmo quando isto nem sequer parece plausível. A teoria clássica seria, em tese, um modelo baseado em pressuposições relativamente fracas. As pressuposições fracas permitem que a teoria seja aplicada numa grande variedade de circunstâncias. Seus conceitos e procedimentos são fáceis de aprender e de usar e, praticamente, não existe nenhuma prova que não se beneficiaria quando analisada sob a prisma do modelo clássico. No entanto, as pressuposições relativamente fracas fatalmente acabam limitando o alcance da teoria clássica, produzindo resultados igualmente limitados.

Em termos operacionais, o ferramental estatístico do modelo clássico inclui:

- a escala somatória simples, definida pela fórmula de escores $X_i = w\sum_{j=1}^{L} x_{ij}$ com o constante w = 1 quando o resultado aparece na forma do total de acertos, ou w = 1 / L, quando o resultado é apresentado na forma da proporção de acertos;
- a média, o desvio padrão e o percentil dos escores dos examinados, que se referem à distribuição dos resultados da prova;
- um coeficiente da fidedignidade do teste, tipicamente, uma aplicação da conhecida fórmula 20 de Kuder & Richardson (1937), de consistência interna, um caso específico do coeficiente alfa de Cronbach (1951);
- o valor-p, ou seja, a proporção dos examinados que acerta a resposta do item, também conhecido como índice de dificuldade do item;
- o coeficiente r_{pbi}, ou seja, a correlação ponto-bisserial entre a resposta correta do item e o escore da prova, também conhecido como índice de discriminação ou índice de fidedignidade do item (pode ser substituído pelo coeficiente de correlação bisserial ou a amplitude inter-quartílica da proporção de acertos D).

Aplica a teoria clássica qualquer pessoa que depende dessas estatísticas para examinar as qualidades técnicas das provas. A abrangência do alcance da teoria clássica está evidente no uso da escala somatória simples por professores em sala de aula para representar o desempenho dos alunos, prática das mais difundidas, que tornou-se quase universal no mundo todo. Na medida em que a escala somatória simples capta o desempenho no conjunto dos itens que compõe a prova, ela é defensável pela sua

simplicidade e por oferecer resultados com níveis de precisão e poder discriminatório satisfatórios na grande maioria dos casos. De fato, investigações do uso de ponderações diferenciadas por item, w_j, particularmente ponderações específicas selecionadas para maximizar a fidedignidade dos escores de prova, revelam que esta aprimoramento é quase sempre desnecessário (Nunnally, 1967, 278). Em quase todos os casos, é mais fácil aumentar a precisão da prova mediante um simples acréscimo de itens do que se preocupar com a contribuição relativa de cada item.

Ao escolher a prova enquanto unidade de análise de preferência, a teoria clássica produziu (e ainda produz) resultados de utilidade imediata comprovada. No entanto, ao priorizar a prova, em vez do item, a teoria clássica não oferece preceitos formais para explicar o comportamento do item, apesar de ser este a matéria prima que compõe as provas. Do cálculo do escore total a partir das respostas de item, percebe-se que o desempenho da prova depende fundamentalmente do desempenho de cada item. Na teoria clássica, portanto, o interesse nas características técnicas dos itens corre em função da intenção de entender as características operacionais da prova como um todo. O conhecimento das características de item permitiria entender e prever as qualidades peculiares de qualquer prova em particular e ainda aprimorar suas características psicométricas para otimizar o desempenho em qualquer situação aplicada. A utilidade prática de cada parâmetro de item depende de sua relação clara, simples e específica com o escore total da prova.

Mesmo assim, caso uma prova contivesse itens que se inter-relacionassem de forma tão orgânica que as características de cada item variasse marcadamente quando apresentado desvinculado dos outros itens, não haveria motivo para deter-se na análise individual dos itens. Da mesma maneira, caso o comportamento do item fosse determinado pelas características singulares de cada grupo de examinados, a análise individual dos itens não teria nenhuma utilidade prática. As duas circunstâncias revelam pressuposições importantes subjacentes à análise de itens. Pressupõe-se que certas características quantificáveis do item permaneçam invariantes, dentro de limites práticos, em todos os diversos situações e contextos aplicados. Desprovido de consistência sob condições variantes, haveria pouca utilidade em analisar as características do item. Por esta razão, procura-se obter parâmetros de item que permaneçam inalterados em provas equivalentes que contêm diferentes conjuntos de itens. Da mesma forma, os parâmetros de item devem permanecer invariantes de um grupo de examinados para outro.

Na falta de uma orientação mais clara a partir da própria teoria clássica, técnicos responsáveis pela elaboração de testes seguem cegamente com a aplicação de estatísticas de itens desenvolvidas originalmente para testes referenciados a normas nacionais. As dificuldades começam logo a surgir, na própria variância dos escores observados. Sendo limitados pelos valores de zero e um, os valores-*p* dos itens não podem formar uma escala linear ou de intervalo. Diferenças iguais nos valores-*p* não representam diferenças iguais na dificuldade dos itens. Ao somar itens de dificuldade arbitrária e desconhecida, com sua variância espremida artificialmente entre zero e um, transfere-se todos os defeitos dos itens para o escore total da prova. Deparamos, assim, numa escala que não é linear e nem sequer previsível. Embora a teoria clássica seja útil quando aplicada a qualquer prova isolada em particular, ela é notoriamente deficiente para orientar uma empreitada científica de maior envergadura.

Vejamos, primeiro, que a proporção de acertos de um único item varia de acordo com o nível de habilidade dos examinados. O mesmo item, quando aplicado a um grupo de habilidade alta e a outro de habilidade modesta, terá valores-*p* diferentes nos dois casos. Freqüentemente, a ordem de grandeza dos valores-*p* de dois itens que medem a mesma habilidade se invertem quando administrados a grupos com habilidades médias diferentes. Por exemplo, o item A pode ter um valor-p mais alto do que o item B quando aplicado a um desses grupos, mas um valor-*p* mais baixo no item B quando administrado ao outro grupo. As discrepâncias assim encontradas são autênticas e não constituem de maneira alguma

aberrações artefatuais ou meros erros amostrais. O comportamento das respostas de item é simplesmente complexo demais para ser representado pelas proporções de valores-*p*.

Em segundo lugar, o poder discriminatório dos itens varia de acordo com a heterogeneidade dos níveis de habilidade da população investigada, a homogeneidade dos itens que compõem o teste e ainda de acordo com a dispersão dos valores-*p* dos itens que compõem o teste. O mesmo item, quando administrado a um grupo de habilidade homogênea e ainda outro grupo, com maiores disparidades de habilidade, produz valores-*d* diferentes nos dois casos. Da mesma forma, um item incluído num teste junto com outros itens de conteúdo e dificuldade homogêneos adquire um valor-*d* diferente do esperado quando o mesmo item integra um teste de conteúdo e dificuldade mais heterogêneos. Os valores-*d* são insuficientes para caracterizar o comportamento de um item e interagem com outros itens ao dependerem do escore da prova. Juntos, valores-*p* e valores-*d* não conseguem representar o desempenho de um item.

As medidas de fidedignidade da prova variam de acordo com o desvio padrão dos escores do teste e em relação aos valores-*p* e valores-*d* dos itens do teste, tudo que depende das habilidades particulares das pessoas examinadas e ainda da composição do teste. Este fato corre como conseqüência natural do efeito da variância sobre o cálculo do coeficiente de correlação. Portanto, um único teste não possui uma única fidedignidade. Seu cálculo é sempre feito em função de um determinado grupo. O mesmo teste, quando aplicado em outras circunstâncias terá coeficientes de fidedignidade proporcionais à variabilidade da distribuição de habilidades. Quando esta variabilidade for grande, o coeficiente será alto; ao contrário, a fidedignidade será baixa. O mesmo instrumento terá um poder discriminatório diferente em cada caso.

A qualidade variável da fidedignidade foi notada por Vianna (1973) ao examinar alunos que respondem a duas formas de um teste. Em seu conjunto, o grupo de alunos exibe uma grande variabilidade de habilidades, conforme mostra os desvios padrão de escores observados num teste relativamente curto. Não é de se estranhar, portanto, ao encontrar um coeficiente de fidedignidade de 0,93 para o grupo como um todo. No entanto, quando se divide os alunos em três grupos de acordo com seu desempenho no teste, calcula-se um coeficiente de fidedignidade de apenas 0,61 para o grupo de menor rendimento, devido à notável homogeneidade interna deste grupo. Ao contrário, juntando os dois grupos extremos, aumentando ainda mais a variabilidade de habilidade, o coeficiente de fidedignidade sobe para 0,95. Nestas circunstâncias, não parece plausível pressupor que um único erro padrão, calculado a partir do coeficiente de fidedignidade do conjunto de medidas, represente o erro em todos os pontos da escala do teste.

Percebe-se que a estatística da teoria clássica tem relevância apenas em situações extremamente limitadas, por exemplo, quando o mesmo item é aplicado á mesma população como parte de um teste rigorosamente paralelo. Essas condições se dão apenas em circunstâncias triviais. Onde a incerteza e a novidade predominam, justamente nas circunstâncias que mais exigem uma orientação teórica, os elementos da teoria clássica pouco ou nada informam sobre os procedimentos a serem adotados na produção de testes considerados adequados do ponto de vista técnico. Quando a população-alvo é diferente do grupo examinado antes, as estatísticas da teoria clássica não se aplicam mais. Evidentemente, essas dependências de grupo comprometem a utilidade prática da teoria clássica. A prova se assemelha a um termômetro que mede a temperatura de uma substância. Na teoria clássica, com as dependências de grupo, é como se os intervalos do termômetro dependesse da temperatura da substância. Fica problemática comparar dois grupos em qualquer sentido significativo quando o a escala do instrumento depende do nível de habilidade.

Na teoria clássica, as medidas de habilidade das pessoas examinadas, apresentadas na forma do número ou da proporção de respostas corretas, dependem da dificuldade dos itens escolhidos para compor o teste. O significado disso se torna evidente quando se tenta

comparar pessoas que respondem a diferentes formas da prova. No contexto do modelo clássico, precisa-se introduzir procedimentos especiais, nem sempre fáceis de elaborar ou de aplicar, para compensar a dificuldade variável dos itens e tornar-se equivalentes os resultados de prova. Na ausência de medidas invariantes do desempenho dos itens, a previsão, especificação e montagem de formas rigorosamente paralelas se tornam muitas vezes impraticáveis, impossibilitando a comparação dos resultados, o que vai de encontro à lógica do empreendimento científico. É como se a temperatura da substância dependesse do termômetro em particular que foi aplicado. A tentativa de comparar dois grupos se torna onerosa quando os instrumentos aplicados produzem medidas incompatíveis entre si.

Constitui-se uma desvantagem bastante séria a dependência da estatística clássica da composição dos grupos investigados e das características dos itens que compõem os testes. Felizmente, hoje existe a possibilidade prática de representar as características dos itens de teste sem depender do grupo examinado e ainda descrever a proficiência das pessoas examinadas sem depender da dificuldade do teste utilizado para este fim. Durante a década de 40, uma nova teoria de mensuração começou a ser formulada a partir das respostas a um único item de teste. O primeiro desses modelos, que usava a ogiva normal para representar as respostas de item, foi desenvolvido por Ferguson (1942), Lawley (1943) e Finney (1944). Esse modelo foi usado pelos psicometristas Brogden (1946) e Tucker (1946) para desenvolver certas considerações teóricas. Seguiu-se uma monografia clássica escrita por Lord (1952) que propunha substituir o desvio normal correspondente ao valor-p como medida da dificuldade do item. No entanto, ao tratar-se das práticas atuais e perspectivas futuras em psicometria, na primeira edição do prestigioso volume Educational Measurement lançado pelo American Council on Education (Lindquist, 1951), a nova teoria sequer mereceu menção. Vieram em seguida contribuições importantes oferecidas por Guttman (1944, 1950, 1954), Lazarsfeld (1959), Rasch (1960), Birnbaum (1968) e Lord & Novick (1968) sem, contudo, alterar a opinião da maioria das pessoas preocupadas com as medidas cognitivas, onde o estudo das respostas de item constituía apenas um campo fértil para especulações teóricas.

No entanto, por volta de 1970, o potencial da nova teoria se pronunciava. Em duas páginas e meia de um capítulo da segunda edição de Educational Measurement (Thorndike, 1971), Angoff afirmou que a grande esperança dos partidários da teoria da resposta ao item é que os constantes de item se demonstrem relativamente invariantes em diversas populações e que a habilidade inferida permaneça relativamente invariante em diferentes conjuntos de itens. Angoff compreendeu com clareza a importância das qualidades de invariância dos novos parâmetros de item e alertou que estes poderiam "provocar grandes inovações nas medidas cognitivas" (p. 529).

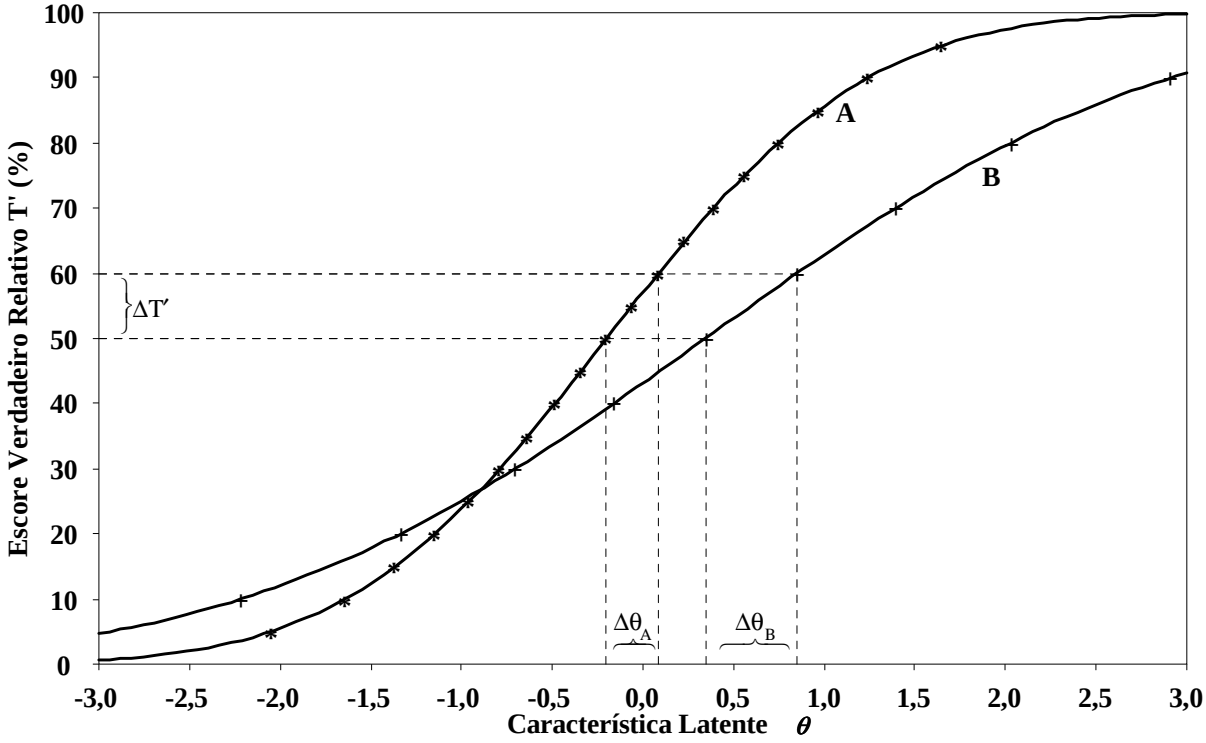
Nas duas décadas seguintes, grande parte da promessa prevista em 1971 acabou se realizando, acompanhado por uma enorme profusão de artigos e livros especializados. Referências às inovações mais recentes podem ser encontradas em trabalhos por Bock (1972), Samejima (1969, 1972, 1973, 1974, 1977) e vários autores europeus, por exemplo, Fischer (1976) e Fischer & Formann (1982). Recentes resumos de literatura aparecem em Traub & Wolfe (1981), Traub & Lam (1985) e Hambleton (1989). Já na terceira edição do volume Educational Measurement lançado em conjunto pelo National Council on Measurement in Education (EUA) e o American Council on Education (Linn, 1989), a teoria da resposta ao item é citada como uma de duas inovações que atualmente exercessem o maior impacto sobre a prática da aplicação de testes e medidas (a segunda inovação seria a tecnologia de informática). Na nova edição, os princípios e aplicações da nova teoria ganharam um enorme relevo em capítulo a parte e ainda em várias seções de outros capítulos que tratam de escalas, normas, procedimentos para tornar resultados de provas equivalentes, investigações do viés de conteúdo nas provas e ainda sobre tecnologias computacionais introduzidas na produção e aplicação de testes. Em todas essas áreas, a nova teoria estabeleceu sua presença, senão sua predominância, mundo afora, ganhando adeptos entre técnicos do setor, editoras de testes, secretários de ensino, superintendentes

de distritos escolares e especialistas em recursos humanos. A velocidade de sua expansão nos países desenvolvidos sugere que, em breve, pode até alcançar o status de paradigma predominante em psicometria.

Uma compreensão do significado desse novo enfoque teórico se obtém ao identificar pressupostos e práticas da teoria de escores verdadeiros que se apresentam como limitações, tautologias ou de alcance insuficiente para as finalidades das medidas em diversas circunstâncias práticas. À luz das insuficiências, pode-se até estranhar saber que a teoria clássica tenha sido útil em algum momento. No entanto, as diferenças que distinguem a teoria da resposta ao item da teoria clássica dizem mais a respeito do nível superior de compreensão alcançado pela primeira do que a qualquer conflito teórico existente entre os dois. A teoria da resposta ao item oferece meios para lidar com problemas que permanecem intratáveis devido às limitações da teoria clássica.

Gráfico 2

Duas Provas A e B: Escalas Curvilineares e Variantes, sem Pontos de Referência Bem Definidos



No centro das deficiências do modelo clássico é o fato de que as unidades de X (e de T) não formam uma escala de intervalo. O incremento de habilidade evidenciado pelo acerto de mais um item na prova A não encontra necessariamente um paralelo no acerto de mais um item na prova B, mesmo quando A e B medem a mesma habilidade e contêm o mesmo número de itens. O próprio conceito da habilidade antecipa a existência de uma escala subjacente (uma característica latente) com unidades de intervalo iguais. Se bem que não se pode demonstrar que a característica latente θ tenha uma escala de intervalos iguais, existe um precedente para exigir essa qualidade em medidas cognitivas, sendo a linearidade uma das qualidades encontradas na grande maioria dos sistemas de medida. Pressupõe-se, também, que θ seja contínuo, o que inclui todos os valores racionais ao longo de sua extensão. Nestas circunstâncias, o escore verdadeiro T e a característica latente θ representam a mesma habilidade, mas em escalas diferentes, com $0 < T < L$, onde L representa o maior escore possível na prova, e $-\infty < \theta < \infty$.

O gráfico de escores verdadeiros relacionados a uma característica latente revela uma a curva que é necessariamente monotônica, não linear e ainda de escala de intervalos desiguais. No Gráfico 2, o escore verdadeiro relativo T', apresentado na forma de uma porcentagem, é relacionado com o valor de θ para testes A e B da mesma característica. A incompatibilidade dos valores verdadeiros nos dois testes é demonstrada de duas maneiras:

1. O valor da coeficiente da característica latente correspondente a *T'* = 50% depende do teste escolhido. A disparidade nos valores correspondentes depende da dificuldade dos itens que compõem os dois testes. Ao obter um valor de θ mais baixo para *T'* = 50% no teste A, conclui-se que o teste A é mais difícil que o B.
2. Um determinado aumento em *T'* no teste A resulta numa mudança na característica θ_A de magnitude diferente da mudança correspondente encontrada na característica θ_B. Os escores verdadeiros são incompatíveis com os princípios de uma escala de intervalo.

As limitações que aparecem quando se usa escores verdadeiros em vez das medidas de habilidade também contaminam o conceito de erro de medida. Sem recorrer ao conceito de característica latente, passa despercebida que um escore relativo de 95% não possui o mesmo grau de precisão que um escore verdadeiro de 50%. O erro padrão da teoria clássica é apresentado nas unidades dos escores observados e de fato diminui assim que T' se aproxima dos limites de 0 e 100%, respectivamente (Lord & Novick, 1968, 385). No entanto, conforme se vê no Gráfico 2, é justamente nesses limites, onde a variação de uma unidade em T' produz o maior aumento na variação de θ, que aumenta a incerteza quanto ao valor real de θ. O problema é superado quando se expressa os erros de medida em termos de θ: quando o valor de θ vai para os extremos, a variância dos erros aumenta. O escore T' = 95% se transforma num valor de θ com erro de precisão maior do que aquele que resulta da transformação de um escore de 50%.

No entanto, talvez o aspecto que serve melhor para mostrar as limitações do modelo clássico gira em torno do que os teóricos das características latentes chamam do "ajuste" dos dados ao modelo de resposta. Quando os itens são ordenados de acordo com seu grau de dificuldade, do mais fácil para o mais difícil, e se ordena as pessoas do menos para o mais hábil, obtém-se uma matriz de dados que confronta dados observados com os resultados previstos pelo modelo. Uma matriz formada das respostas de cinco pessoas a três itens de teste serve para exemplificar a noção de ajuste. Na Tabela I, relaciona-se as pessoas de acordo com o valor dos escores observados e os itens segundo a conta do número de pessoas que os acertam.

Tabela 1

Uma Configuração de Respostas Guttman, Imperfeita

Pessoas Ordenadas segundo a Habilidade	Itens Ordenados Segundo a Dificuldade					Escore da Pessoa
	1	***2***	***3***	***4***	***5***	
1	1	1	0	0	0	2
2	1	0	1	1	0	3
3	1	1	1	0	1	4
Conta de Itens	3	2	2	1	1	

A configuração de respostas forma uma escala Guttman quando todos os acertos se encontram no diagonal inferior à esquerda dessa matriz, formando um triângulo. A Tabela I apresenta duas inversões à configuração triangular idealizada por Guttman (1944, 1950, 1954). Estas ocorrem quando um acerto, representado por 1, aparece à direita de um desacerto, representado por 0. Quando os dados se conformam ao modelo Guttman, o

triângulo é perfeito. A característica definitiva de uma escala Guttman é que o conhecimento do escore observado explica tudo sobre o comportamento da pessoa em cada item, ou seja, tudo que se pode saber a respeito da habilidade da pessoa. Nessas circunstâncias ideais, o conhecimento do escore observado é suficiente para determinar as respostas a todos os itens. Isto se torna um critério para julgar a adequação do modelo Guttman numa instância específica qualquer.

Ao antecipar por vários anos o desenvolvimento dos modelos probabilísticos de resposta ao item, Guttman apresentou esta escala determinista com seu conceito de reprodutibilidade. Ele postulou uma característica unidimensional latente subjacente a seu modelo de resposta e desenvolveu um procedimento capaz de relacionar pessoas e itens juntos numa única dimensão. Mas sua formulação não probabilística preveniu uma análise satisfatória da noção de ajuste. A natureza determinista do modelo rejeitou qualquer possibilidade de erros de medida, o que acabou se demonstrando pouco realista. Um conceito mais adequado de ajuste emergiu apenas no contexto das configurações de resposta caracterizadas por probabilidades, o que permitiu uma consideração explícita dos erros de medida aleatórios. No entanto, em sua essência, os teóricos das características latentes compartilham a mesma preocupação de Guttman ao relacionar explicitamente as configurações de resposta a um modelo. Ao examinar a consistência do modelo com as configurações de resposta, a teoria das características latentes vai muito além dos princípios que orientam o uso das escalas somatórias simples (a soma dos acertos ou sua normalização) na teoria clássica.

Ao relacionar escores verdadeiros com coeficientes de características latentes, aproxima-se de uma definição. Na teoria clássica, tratam-se os resultados de prova como se fossem amostras do comportamento numa determinada área de atividade, examinados para fazer previsões ou tirar inferências generalizáveis. O comportamento observado é considerado um sinal de outros comportamentos relacionados que, em seu conjunto, formam uma única classe de respostas comum. As respostas mudam todos na mesma direção em função de estímulos comuns presentes no ambiente. É suficiente captar apenas uma pequena amostra dessa mudança de estado, concebida como uma condição temporária de sentimento ou de mentalidade, nível transitório de excitação ou atividade presente provocada, para inferir a presença de outros comportamentos generalizáveis. Os comportamentos se interrelacionam consistentemente porque são solicitados e mantidos em função das mesmas contingências ambientais, organizados pelos mesmos processos. Veja-se, portanto, que existe uma certa afinidade entre a teoria clássica, em termos dos princípios de medidas, e a perspectiva comportamentalista, enquanto teoria psicológica.

Antes pelo contrário, a teoria de características latentes interpreta os resultados de teste como sinais de profundos processos ou estruturas psicológicas subjacentes. Apesar de considerável amplitude na variabilidade das condições, situações e circunstâncias, estas se revelam características, atributos ou disposições bastante estáveis, capazes de se manifestarem ou de provocar comportamentos em qualquer momento relevante. Ao especificar o que acontece quando uma pessoa responde a um teste, pressupõe-se que as regularidades de resposta podem ser explicadas por certas características subjacentes. Por não serem características observáveis, passíveis de medição direta, são muitas vezes conhecidas como características ou habilidades latentes. A característica seria uma abstração conceitual que pode ser manipulada, limitada, variada ou de outra forma transformada conforme os objetivos previstos para seu uso. Seu valor depende exclusivamente de sua utilidade como ferramenta na compreensão da experiência revelada pelo comportamento. Uma característica latente seria, portanto, uma construção psicológica cuja presença é postulada para explicar a variabilidade replicável das observações de comportamento encontrada em certas situações específicas bem definidas. A questão da existência física das características psicológicas não está em jogo. É suficiente que a pessoa se comporte como se tivesse determinada característica e que esta característica determinasse substancialmente seu comportamento.

Se bem que a teoria de características latentes tem a estimação da habilidade como seu objetivo, o modelo de respostas constitui o cerne da teoria. Esse modelo especifica a relação entre o desempenho do examinado num teste e as características ou habilidades não observáveis que se pressupõem subjacentes ao instrumento. A relação entre as quantidades observáveis e não observáveis é representada por uma função matemática ou modelo matemático que representa a probabilidade (e não a certeza) da ocorrência de um evento previsto. Na medida em que pode ser demonstrado que as representações do modelo se aproximam da realidade observada, suas funções podem ser usadas para qualificar as características de indivíduos e ainda para prever ou explicar seu comportamento nesta base.

Os modelos de resposta se baseiam em três conjuntos de elementos: uma enumeração dos eventos possíveis, uma identificação de parâmetros latentes considerados necessários para representar o desempenho sistemático dos itens, e uma identificação dos parâmetros latentes considerados necessários para representar o comportamento sistemático das pessoas. A formulação, ao se tratar de elementos de natureza especificamente sistemática, faz referência a parâmetros identificados como necessários, relacionados com características que são persistentemente replicáveis ou invariantes dentro de um quadro de referência bem definida. Não se nega a existência de outras características que poderiam condicionar as respostas -- de certo, deveriam existir uma infinidade delas -- mas não haveria motivo para sua representação a não ser que se refiram a estas mesmas qualidades. Este ponto ganha relevância ao apreciar as diversas alternativas de modelo existentes. Cada modelo apresenta seu próprio conjunto de parâmetros como ingredientes e assim afirma que as outras possibilidades podem ser ignoradas ou então controladas.

A apresentação a seguir obedece a certas restrições para enfatizar o essencial. Assim, em primeira instância, trata-se, preferencialmente, dos modelos que limitam os eventos possíveis ao conjunto binário de 1 e 0, que normalmente representa acertos e desacertos, respectivamente. Este formato é apropriado para itens de múltipla escolha, uma vez que os resultados sejam recodificados para representar acerto e erro. É conveniente, também, considerar que o comportamento da pessoa seja explicado por apenas uma característica latente, representada por um único parâmetro de pessoa denominado θ. A característica poderia ser uma capacidade cognitiva, uma medida de desempenho ou de rendimento, uma competência básica ou habilidade, uma variável de personalidade ou uma atitude. Embora uma diversidade de outras influências possam interferir no teste, pressupõe-se que um componente ou fator dominante, medido pelo teste, predomina no resultado. Finalmente, a forma matemática do modelo é limitada a uma função que aumenta monotonicamente com a habilidade latente θ, ou seja, que a probabilidade de um acerto aumente junto com θ, o que possibilita a mensuração.

Gráfico 3

Curva de Característica do Item *j*

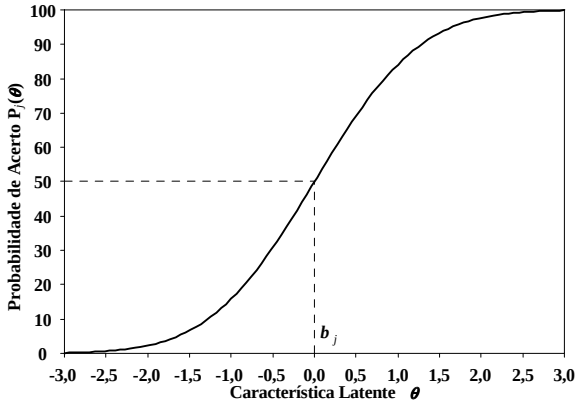
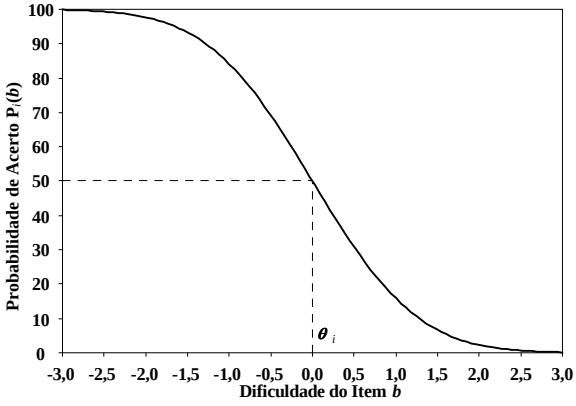


Gráfico 4

Curva de Característica da Pessoa *i*



Tendo em vista que o conhecimento das características latentes se obtém apenas a partir das respostas de item, é costumeiro formular o modelo matemático em termos da chamada "curva de característica do item". A curva de característica do item é uma função matemática que relaciona a probabilidade de um acerto na resposta do item a um valor especifico de θ encontrado em algum ponto ao longo da escala da habilidade latente. Esta nada mais é que uma função não-linear para a regressão do escore de item na característica ou habilidade medida pelo teste. Uma formulação desse tipo se encontra no Gráfico 3, onde a probabilidade de acerto $Pj(\theta)$ aparece como função da variável independente θ, com a dificuldade do item *j* fixa. Para qualquer item, a probabilidade de sucesso aumenta de acordo com a habilidade da pessoa. A forma da curva, sua localização em relação à abcissa e outras características dependem da função e dos parâmetros do item j. Ao postular uma relação probabilística explícita, $Pj(\theta)$ se aproxima de seus limites 0 e 1 assintoticamente, o que dá a forma sigmóide da curva de característica do item.

Pode-se visualizar essa mesma relação de uma outra perspectiva. Supondo que se mantenha a habilidade da pessoa constante ao variar a característica latente do item, surge a "curva de característica da pessoa" encontrada no Gráfico 4. A nova curva representa o desempenho esperado de uma pessoa específica ao enfrentar itens de dificuldade variável. A probabilidade de acerto $Pi(b)$ seria uma função da variável independente b, que representa a dificuldade do item j, ao manter-se constante a habilidade da pessoa *i*. Para qualquer pessoa, a probabilidade de sucesso diminui de acordo com a dificuldade do item.

Nos modelos de resposta ao item, as probabilidades $Pj(\theta)$ e $Pi(\theta)$ têm significados nítidos em termos de processos. Na teoria clássica, o processo visualizado envolve a administração do mesmo item à mesma pessoa um número infinito de vezes, o que resulta no escore verdadeiro *xj*. Às vezes se dá essa mesma interpretação a $Pj(\theta)$. Neste caso, $Pj(\theta)$ seria o resultado esperado, tendo em vista a distribuição das propensidades de resposta. No entanto, proponentes dos modelos de resposta ao item preferem duas interpretações alternativas. A primeira visualiza um acervo de itens com dificuldades e outras características técnicas iguais. Para um item escolhido aleatoriamente desse acervo, $Pj(\theta)$ seria a probabilidade de uma resposta correta de um examinado com nível da habilidade θ. A segunda interpretação envolve uma amostra de pessoas, todas da mesma habilidade. $Pj(\theta)$ seria a probabilidade de um acerto por um examinado de habilidade θ, quando a pessoa for selecionada aleatoriamente desse grupo. O termo "teoria de características latentes" refere-se à probabilidade de um acerto do ponto de vista da pessoa, enquanto a expressão "teoria da resposta ao item" enfatiza o mesmo comportamento a partir da

perspectiva do item. O termo "teoria das curvas de característica" faz jus às duas perspectivas mas é menos sugestivo quanto aos objetivos.

A identificação de θ e *b* na mesma escala que corre entre -∞ e ∞ é uma contribuição importante da teoria de características latentes, onde pessoas e estímulos compartilham o mesmo status conceitual. θ e *b* representam a localização das pessoas e dos itens com respeito à variável latente que compartilham. Isto permite que se relacione itens e pessoas na mesma escala, usando a mesma métrica. Costuma-se dizer que θ representa a posição da pessoa e *b* o valor de escala do item. Os parâmetros de pessoa e de item assumem seu papel no modelo matemático ao especificar a probabilidade da experiência conjunta, quando a pessoa *i* responde ao item j.

Dos diversos modelos de resposta, o do matemático dinamarquês Georg Rasch (1901-1980) é o mais parcimonioso, reflexo de uma lógica sucinta e elegante (Wright, 1977; Andrich, 1988). Rasch (1960) critica implicitamente a teoria dos escores verdadeiros ao reclamar que, como estudante, sempre foi avaliado num conjunto de tarefas de dificuldade arbitrária e sempre comparado a uma população de referência indefinida. Ele defende medidas de habilidade que não se associam a essas relatividades e, ao contrário, proporcionem parâmetros de pessoa compreendidos no sentido absoluto. O raciocínio matemático desenvolvido por Rasch toma como objetivo separar a habilidade da pessoa da dificuldade do item. Ao alcançar esta separação, as características dos itens podem ser estimadas sem depender da habilidade das pessoas e as habilidades sem depender da dificuldade dos itens. Ganha-se com isso parâmetros de item e de pessoa que seriam invariantes numa grande variedade de circunstâncias, o que assegura sua objetividade.

Rasch define as condições necessárias para as medidas objetivas nos seguintes termos:

> *A comparação entre dois estímulos deve ser independente de quaisquer pessoas em particular que serviram para fazer a comparação; e também deve ser independente de quaisquer outros estímulos da classe em consideração que também foram ou poderiam ser comparadas.*
> *De maneira simétrica, a comparação entre dois indivíduos deve ser independente de quaisquer estímulos em particular da classe em consideração que serviram para fazer a comparação; e também deve ser independente de quaisquer outros indivíduos que também foram comparados, nesta ou em alguma outra ocasião [Rasch, 1961: 322].*

Rasch dá ênfase à separação dos parâmetros θ e b como motivo para o desenvolvimento de seu modelo. Sua formulação define equivalências matemáticas formais para cada um dos princípios citados acima.

O procedimento para uma "análise de itens livres de considerações de amostra" (Wright & Panchapakesan, 1969) baseia-se numa simples noção do que acontece quando uma pessoa enfrenta qualquer item de teste. O modelo prevê que o resultado dessa experiência conjunta seja um produto determinado pela habilidade da pessoa e a dificuldade do item. Tanto mais hábil a pessoa, tanto melhor seriam suas chances de acertar qualquer item. Tanto mais difícil o item, tanto menos provável que qualquer pessoa consiga solucioná-lo. As chances de êxito são altos para candidatos hábeis em itens fáceis e baixas para candidatos sem qualificação em itens difíceis. A vantagem da pessoa i no item j seria, portanto, igual à razão de sua habilidade θi* relativa à dificuldade do item bj*, isto é, θi* / bj*, quando se define θi* e bj* numa escala [0, ∞].

A probabilidade da ocorrência de um evento qualquer relativa a sua não ocorrência é definido pela razão de verossimilhança de P / (1 - P), onde P é a probabilidade de uma ocorrência qualquer. No contexto do confronto entre o examinado e o item, a razão de verossimilhança toma a forma da razão de Pj(θi) para 1 - Pj(θi), onde Pj(θi) é a probabilidade da pessoa i acertar o item j, Pr{xij = 1}. Nesta base, equivalem-se as chances de um acerto relativo a um desacerto com a habilidade da pessoa relativa à dificuldade do item:

$$\frac{\Pr\{x_{ij}=1\}}{\Pr\{x_{ij}=0\}} \equiv \frac{\theta_i}{b_j} = \frac{P_j(\theta_i)}{1-P_j(\theta_i)}, \tag{11}$$

a partir da qual pode-se inferir que a probabilidade de um acerto é:

$$\Pr\{x_{ij}=1\} \equiv P_j(\theta_i) = \frac{\theta_i^*}{\theta_i^* + b_j^*} = \frac{\theta_i^* / b_j^*}{1+\theta_i^* / b_j^*} \tag{12}$$

e a probabilidade de um desacerto:

$$\Pr\{x_{ij}=0\} \equiv 1-P_j(\theta_i) = 1-\frac{\theta_i^*}{\theta_i^* + b_j^*} = \frac{b_j^*}{\theta_i^* + b_j^*} = \frac{1}{1+\theta_i^* / b_j^*}. \tag{13}$$

O fator de normalização gij = 1 + θi* / bj* assegura que as probabilidades somam a um, ou seja, Pr{xij = 1} + Pr{xij = 0} = 1.0. Também, com θi* > 0 e bj* > 0, o resultado se mantém entre os limites previstos para uma probabilidade, ou seja, 0 < Pr{xij = 1} < 1. Para tomar um exemplo, caso existissem habilidade θi* = 5 e dificuldade bj* = 2, a ocorrência de uma resposta de xij = 1 em vez de xij = 0 seria representada pela razão de verossimilhança de 5 : 2. No entanto, a simples probabilidade da ocorrência xij = 1 seria 5 / (5 + 2) = (5 / 2) / (1 + 5 / 2) = 0,714. Ao fixar o valor de bj* e variar o valor de θi*, a fórmula em (12) traceja a curva de característica de item encontrada no Gráfico 3. Ao fixar θi* e variar bj*, ela traceja a curva de característica de pessoa do Gráfico 4.

Embora as formulações em (11)-(13) sejam mais claras do ponto de vista conceitual e apareçam no trabalho publicado por Rasch, elas geralmente aparecem na literatura estatística de outra forma. Se partíssemos da pressuposta de que:

$$e^{\lambda} = \frac{P_j(\theta_i)}{1-P_j(\theta_i)}, \tag{14}$$

onde e = 2.718... é a base dos logaritmos naturais, um pouco de manipulação algébrica demonstra que a probabilidade de um acerto é definida pela função cumulativa da ogiva logística, Ψ(λ):

$$\Pr\{x_{ij}=1\} \equiv P_j(\theta_i) \equiv \Psi(\lambda) = \frac{e^{\lambda}}{1+e^{\lambda}} = \frac{1}{1+e^{-\lambda}}, \tag{15}$$

e a probabilidade de um desacerto:

$$\Pr\{x_{ij}=0\} \equiv 1-P_j(\theta_i) \equiv 1-\Psi(\lambda) = 1-\frac{e^{\lambda}}{1+e^{\lambda}} = \frac{1}{1+e^{\lambda}}. \tag{16}$$

Pode-se, então, escrever:

$$\theta_i^* = e^{\theta_i} \quad \text{e} \quad b_j^* = e^{b_j} \tag{17}$$

para designar respectivamente um novo parâmetro de habilidade θ_i e um novo parâmetro de dificuldade b_j, cada um definido numa escala transformada. A partir daí, obtém-se a formulação mais convencional. A probabilidade de um *acerto* é:

$$\Pr\{x_{ij}=1\} \equiv P_j(\theta_i) \equiv \Psi(\theta_i - b_j) = \frac{e^{(\theta_i - b_j)}}{1+e^{(\theta_i - b_j)}} = \frac{1}{1+e^{-(\theta_i - b_j)}}, \tag{18}$$

que, em vez de comparar habilidade e dificuldade na base de sua *razão*, faz a comparação através da *diferença* de $\lambda_{ij} = \theta_i - b_j$. Essa diferença determina a probabilidade do que ocorre quando a pessoa *i* aplica sua habilidade para resolver a dificuldade do item *j*. Às vezes, para enfatizar a experiência conjunta relacionando pessoas e itens, a probabilidade aparece na forma de $P(\lambda_{ij})$, onde λ_{ij} seria uma função tanto dos parâmetros de pessoa quanto de item. O modelo Rasch toma a forma de uma ogiva logística, $\Psi(\lambda_{ij})$, baseada na *unidade de probabilidade logística* ou *logit* de $\lambda_{ij} = \theta_i - b_j$, usando terminologia apresentada por Berkson (1944). A *habilidade* da pessoa em logits é o logaritmo natural da razão de verossimilhança relacionando a probabilidade de um acerto com a de um desacerto em itens de dificuldade

encontrada na origem da escala, em zero. A probabilidade de um acerto num item com dificuldade $b_j = 0$ é $P_j(\theta_i) = e^{\theta_i} / (1 + e^{\theta_i})$, a partir do qual a razão de verossimilhança de um acerto é $P_j(\theta_i) / (1 - P_j(\theta_i)) = e^{\theta_i}$, cujo logaritmo é θ_i. A *dificuldade* em logits é o logaritmo da razão de verosimilhança relacionando a probabilidade de um desacerto com a de um acerto por pessoas de habilidade na origem da escala. A probabilidade de êxito de pessoas com habilidade $\theta_i = 0$ num item com dificuldade b_j é $P_j(\theta_i) = e^{-b} / (1 + e^{-b})$, a partir do qual a razão de verossimilhança de um desacerto é $(1 - P_j(\theta_i)) / P_j(\theta_i) = e^{b_j}$, cujo logaritmo é b_j. Logo, a contrapartida da equação (18) seria a probabilidade de um *desacerto*:

$$\Pr\{x_{ij} = 0\} \equiv 1 - P_j(\theta_i) \equiv 1 - \Psi(\theta_i - b_j) = 1 - \frac{e^{(\theta_i - b_j)}}{1 + e^{(\theta_i - b_j)}} = \frac{1}{1 + e^{(\theta_i - b_j)}}. \qquad (19)$$

Tomando proveito do fato de que x seja uma variável binária e da convenção onde xij = 1 represente um acerto e xij = 0 um desacerto, pode-se combinar (18) e (19) numa única fórmula:

$$\Pr\{x_{ij} = x\} \equiv P_j(\theta_i)^{x_{ij}} Q_j(\theta_i)^{1 - x_{ij}} = \frac{e^{(\theta_i - b_j) x_{ij}}}{1 + e^{(\theta_i - b_j)}}, \qquad (20)$$

onde $Q_j(\theta_i) = 1 - P_j(\theta_i)$. Quando $x_{ij} = 1$, o termo $Q_j(\theta_i)^{1 - x_{ij}}$ em (20) se torna uma identidade, sem efeito. Quando, por sua vez, $x_{ij} = 0$, então $P_j(\theta_i)^{x_{ij}}$ se torna uma identidade.

Para aplicar o modelo Rasch, um conjunto de itens de prova é elaborado para aplicação em determinado grupo de pessoas. A partir das respostas de uma amostra de pessoas geralmente muito maior em tamanho do que o número de itens, as propriedades técnicas dos itens são avaliadas. No processo, alguns dos itens podem sofrer modificações, reciclagem e re-avaliação. Essa etapa é conhecida pelo termo de calibração dos itens (Wright & Panchapakesan, 1969; Wright, 1977). Após a calibração dos itens, as posições das pessoas que respondem aos itens nesta ou qualquer outra ocasião podem ser estimadas. O termo medida de pessoa se refere a essa nova etapa. O objetivo procurado através da aplicação do modelo Rasch é de estimar parâmetros de itens e de pessoa que reproduzem com precisão as probabilidades de acerto de cada pessoa a cada item.

Ao recomendar a aplicação de seu modelo, Rasch (1961) pressupõe que 1) as probabilidades de acerto podem ser representados pela função logística, conforme especificada, e 2) pessoas e itens podem ser considerados eventos independentes. A primeira pressuposição tem diversas implicações. Sem dúvida, a implicação mais importante é que os itens representem uma característica latente unidimensional. Nesta hipótese, toda a covariância entre os itens se atribui à variação das pessoas na característica latente sendo medida. Ao agrupar as pessoas segundo a habilidade, conforme constatada no teste, não haveria nenhuma correlação restante entre os itens. A unidimensionalidade implica que cada item ordena as pessoas do mesmo jeito, aumentando a probabilidade de acerto do item de acordo com o escore total da prova.

A segunda pressuposição se refere às condições experimentais da prova. A independência das pessoas significa que as respostas de qualquer pessoa não afetam as respostas de qualquer outra pessoa. Da mesma maneira, a independência dos itens significa que a resposta de uma pessoa a qualquer item não influi sua resposta a qualquer outro item. Dessa maneira, as probabilidades de acerto não variam do conjunto da prova para qualquer subconjunto dos itens. Uma formulação matemática equivalente para o princípio da independência afirma que a probabilidade da ocorrência de um conjunto de eventos é simplesmente o produto das probabilidades de ocorrência de cada um deles isoladamente. A independência requer que quaisquer dois itens permaneçam sem correlação em grupos de pessoas da mesma habilidade. Veja-se, portanto, que a pressuposição de independência segue como conseqüência da unidimensionalidade. A rigor, esse princípio não representa nenhuma pressuposição adicional.

Para considerar os métodos de estimação dos parâmetros θi e bj, é conveniente introduzir $g_{ij} = 1 + \theta_i^* / b_j^*$ para representar o fator de normalização no denominador de (12) e o equivalente $\gamma_{ij} = 1 + e^{(\theta_i - b_j)}$ para a normalização em (18). Sob condições de independência estatística, o modelo Rasch representa a resposta de duas pessoas *i* = 1 e *i* = 2 a um único item *j* nos seguintes termos:

$$\Pr\{x_{1j} = 1;\, x_{2j} = 1\} \equiv \Pr\{x_{1j} = 1\} \Pr\{x_{2j} = 1\}$$
$$= \frac{(\theta_1^* / b_j^*)(\theta_2^* / b_j^*)}{g_{1j} g_{2j}}. \qquad (21)$$

De forma análoga, são independentes as respostas de uma única pessoa *i* a dois itens *j* = 1 e *j* = 2 quando:

$$\Pr\{x_{i1} = 1;\, x_{i2} = 1\} \equiv \Pr\{x_{i1} = 1\} \Pr\{x_{i2} = 1\}$$
$$= \frac{(\theta_i^* / b_1^*)(\theta_i^* / b_2^*)}{g_{i1} g_{i2}}. \qquad (22)$$

A seguir, examina-se em maior detalhe o caso de uma pessoa que responde a dois itens.

Tabela II

Configuração de Respostas do Modelo Rasch

Escore	**Freq.**	**Item *j* Ordenado segundo a Dificuldade**						
r_k	n_k	***1***	***2***	***3***	***4***	***5***	***6***	***7***
0	130	0 *0%*	0 *0%*	0 *0%*	0 *0%*	0 *0%*	0 *0%*	0 *0%*
1	302	110 *36%*	76 *25%*	39 *13%*	31 *10%*	14 *5%*	17 *6%*	15 *5%*
2	439	273 *62%*	254 *58%*	131 *30%*	99 *23%*	39 *9%*	39 *9%*	43 *10%*
3	482	401 *83%*	381 *79%*	250 *52%*	160 *33%*	103 *21%*	89 *18%*	62 *13%*
4	443	385 *87%*	401 *91%*	298 *67%*	250 *56%*	163 *37%*	155 *35%*	120 *27%*
5	294	275 *94%*	283 *96%*	255 *87%*	209 *71%*	175 *60%*	157 *53%*	116 *39%*
6	131	129 *98%*	131 *100%*	123 *94%*	118 *90%*	110 *84%*	93 *71%*	82 *63%*
7	32	32 *100%*	32 *100%*	32 *100%*	32 *100%*	32 *100%*	32 *100%*	32 *100%*
Conta de Item s_j		1605	1558	1128	899	636	582	470

Habilidade	
θ_k	**se(θ_k)**
-∞	--
-2,271	0,621
-1,186	0,582
-0,346	0,639
0,421	0,753
1,213	0,938
2,223	1,329
∞	--

Ajuste do conjunto é χ_j^2 = 2,380, com 1 grau de liberdade e probabilidade $\Pr\{\chi_j^2\}$ = 0,123.

Dificuldade dos Itens	b_j	-1,709	-1,561	-0,389	0,195	0,920	1,085	1,459
	se(b_j)	0,057	0,056	0,050	0,051	0,055	0,056	0,060
	χ_j^2	1,585	3,594	0,788	1,633	1,048	1,373	4,256
	$\Pr\{\chi_j^2\}$	0,208	0,058	0,378	0,201	0,306	0,241	0,039

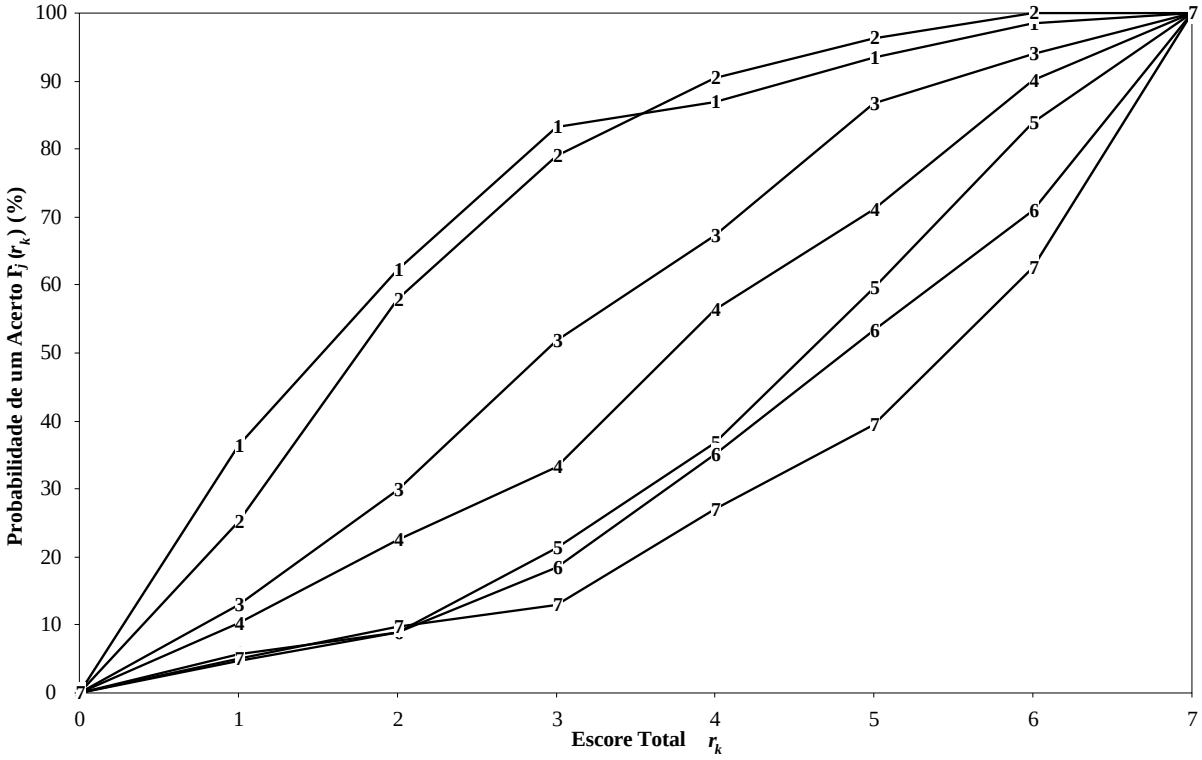
Fonte: Sete itens selecionados de uma prova de direito elaborada pela Fundação Carlos Chagas.

Considere inicialmente a "prova" de apenas sete itens resumida na Tabela II. O primeiro bloco da tabela apresenta o número e a porcentagem de acertos para cada item segundo o escore total da prova. As colunas dos itens *j = 1 .. 7* se encontram ordenadas segundo a conta de respostas corretas do item, *sj*, e as linhas das pessoas são ordenadas de acordo com o escore total recebido na prova, *rk*. Existem oito escores possíveis na prova, o que dá a seqüência *k = 0 .. 8*.

Uma vez que se ordene as colunas e linhas dessa maneira, a tabela apresenta um quadro de referência em duas dimensões, com as maiores probabilidades de acerto encontradas no triângulo do canto inferior a esquerda. Esse esquema faz lembrar a configuração de respostas Guttman. Ao passar da esquerda para a direita em qualquer linha do bloco, as probabilidades de acerto tendem a cair. Em qualquer coluna, as probabilidades tendem a aumentar. À maneira da configuração Guttman, as maiores probabilidades de acerto se encontram no triângulo inferior de esquerda da Tabela.

Gráfico 5

Proporção de Pessoas $P_j(r_k)$ com Escore Total r_k que Acertam o Item *j*



A estrutura assimétrica das porcentagens fica patente logo em seguida, no Gráfico 5, mesmo quando os escores observados rk não formem uma escala linear. As curvas do gráfico contrastem o desempenho de itens fáceis e itens difíceis. As probabilidades de acerto dos itens fáceis, mais à esquerda, sobem mais rapidamente nos escores intermediários do que as probabilidades de acerto dos itens mais difíceis, encontrados à sua direita. Afastando-se dos escores extremos de 0 ou 7, as probabilidades de acerto se distanciam entre si, o que destaca o desempenho diferenciado dos itens. Veja-se que nenhuma função matemática serve para representar a família de curvas encontrada no Gráfico 5. A assimetria encontrada nessa estrutura serve para a estimar os parâmetros do modelo Rasch, o que permite *linearizar* a escala da abcissa.

Tabela III

Razão de Verossimilhança dos Itens 1 e 5 nos Escores r_k = 1 .. 6 da Prova de Direito de Sete Itens

Escore	Freqüência Observada			Freqüência Esperada		
Total r_k	(1,0)	(0,1)	Razão	(1,0)	(0,1)	Razão
1	110	14	7,86	112,36	11,64	9,65
2	257	23	11,17	253,71	26,29	9,65
3	328	30	10,93	324,39	33,61	9,65
4	249	27	9,22	250,09	25,91	9,65
5	116	16	7,25	119,61	12,39	9,65
6	21	2	10,50	20,84	2,16	9,65
Conta de Item s_j	1081	112	9,65	1081,00	112,00	9,65

Teste χ^2 para tabelas de contingência é χ^2 = 2,631, gl = 5 e p = 0,756.

Ao examinar essa estrutura de assimetria, consideramos inicialmente apenas dois itens, por exemplo, os itens 1 e 5 dessa prova. Quando as respostas dos dois itens são iguais, o que gera as configurações de resposta (0,0) e (1,1), nada pode-se inferir sobre a dificuldade relativa desses dois itens. Isto é típico da situação encontrada quando as pessoas recebem um escore de zero ou o máximo possível na prova. Nestes dois casos, ou eles erram os dois itens ou acertam os dois. O modelo Rasch ignora estes casos extremos porque os escores não permitem contrastes entre os itens. Já as configurações (1,0) e (0,1) afirmam que um item é mais fácil (ou mais difícil) do que o outro item, informação que permite estimar os parâmetros de item. Compara-se, então, a probabilidade da ocorrência de um evento (1,0) com a do evento (0,1), utilizando a razão de verossimilhança:

$$\frac{\Pr\{(1,0)\}}{\Pr\{(0,1)\}} = \left(\frac{\theta_i^* / b_1^*}{\theta_i^* / b_2^*}\right)\left(\frac{g_{n1}g_{n2}}{g_{n1}g_{n2}}\right) = \frac{b_2^*}{b_1^*},$$

$$\text{ou seja,} \quad \frac{\Pr\{(1,0)\}}{\Pr\{(0,1)\}} = \frac{1/b_1^*}{1/b_2^*} = \frac{b_2^*}{b_1^*}, \tag{23}$$

onde 1 / *b** seria a *facilidade* do item. Logo, percebe-se que o resultado b_2^* / b_1^* não depende mais do parâmetro de pessoa θ_i^*! A razão das probabilidades desses dois itens é a mesma para *todas* as pessoas e não depende da posição particular de θ_i^*, que varia de pessoa para pessoa. Portanto, as configurações de resposta (1,0) e (0,1) para diferentes pessoas podem ser considerações *replicações* e usadas para estimar a razão de verossimilhança representada por b_2^* / b_1^*. Encontramos, assim, a primeira instância do princípio da *separação*, onde se estabelece a dificuldade do item independente das habilidades das pessoas examinadas ao fazer esta inferência, um princípio demonstrado melhor com a simplicidade e elegância do modelo Rasch, mas, em princípio, uma qualidade compartilhada por todos os modelos de resposta ao item.

Para calcular a razão de verossimilhança dos dois itens do exemplo, precisamos das freqüências das duas configurações de resposta (1,0) e (0,1) discriminados segundo o escore total na prova r_k. Conforme as freqüências apresentadas na Tabela III, estima-se a razão de verossimilhança:

$$\frac{\Pr\{(1,0)\}}{\Pr\{(0,1)\}} = \hat{b}_5^* / \hat{b}_1^* = 1081 / 112 = 9{,}652.$$

Ao lembrar em (12) que Pr{x_5 = 1} aumenta de acordo com o *inverso* de b_5*, conclui-se que o item 5 é mais de nove vezes mais difícil que o item 1. Em outras palavras, caso uma pessoa acerta apenas *um* desses dois itens, é quase *dez vezes* mais provável que o item acertado seja o item 1. A unidade de medida da comparação é a razão de verossimilhança da probabilidade de acertar um desses itens e não o outro quando apenas um deles for acertado.

Para completar, em logaritmos naturais,

$$\log\left(\frac{\Pr\{(1,0)\}}{\Pr\{(0,1)\}}\right) = \log(\hat{b}_5^* / \hat{b}_1^*) = \log(1081/112) = 2{,}267, \quad \text{ou seja,} \quad \hat{b}_5 - \hat{b}_1 = 2.267,$$

o que representa a diferença entre os valores de escala dos dois itens. A unidade de medida da comparação é o *log* da razão de verossimilhança da probabilidade de acertar um desses itens e não o outro quando apenas um deles é acertado.

As estimativas obtidas representam apenas a *razão* de $\hat{b}_5^* / \hat{b}_1^*$ ou a *diferença* $\hat{b}_5 - \hat{b}_1$. Nenhum valor específico pode ser atribuído a qualquer dos dois parâmetros de item. Nestas circunstâncias, é comum impor a restrição de $\hat{b}_1^* \cdot \hat{b}_5^* = 1$ ou o equivalente $\hat{b}_1 + \hat{b}_5 = 0$. Com $\hat{b}_1^* \cdot \hat{b}_5^* = 1$ e $\hat{b}_5^* / \hat{b}_1^*$ = 9,652, segue que $\hat{b}_5^{*2} = 9{,}652$, o que dá $\hat{b}_5^* = 3{,}107$ e $\hat{b}_1^* = 0{,}322$. Na métrica logarítmica, distribuem-se $\hat{b}_1 = -1{,}133$ e $\hat{b}_5 = 1{,}133$ em torno de zero, fixando suas posições de escala.

A comparação numérica serviu para fixar a posição dos itens em termos de uma razão ou de uma diferença. A comparação é independente dos níveis de habilidade das pessoas e, portanto, deve permanecer invariante ao longo da escala de habilidade. Por analogia, as posições de quaisquer duas pessoas podem ser comparadas na base das respostas a vários itens, sem depender dos valores de escala desses itens. Em vez de usar as contas de acerto de item sj para comparar itens, compara-se o total de pessoas em cada escore rk para determinar as posições de habilidade relativa. Ao examinar o caso de duas pessoas de escores diferentes que acertam um único item, prossegue-se a verificar que a razão de habilidades, calculada a partir das freqüências de acerto, não depende dos parâmetros de item bj*. As configurações de resposta para diferentes itens podem ser consideradas replicações e usadas para estimar a razão de verossimilhança entre duas pessoas θ_1^* / θ_2^*. Impõem-se as mesmas restrições de $\hat{\theta}_1^* \cdot \hat{\theta}_2^* = 1$ ou o equivalente $\hat{\theta}_1 + \hat{\theta}_2 = 0$ para fixar as posições de escala. Encontramos, assim, a segunda instância do princípio de separação, onde se estabelece as habilidades das pessoas independentes dos itens em particular que serviram de base para fazer estas inferências, um princípio compartilhado por todos os modelos de resposta ao item.

Enfatizamos a invariância das comparações entre pares de itens e pares de pessoas porque esta constitui um princípio fundamental em todos os modelos de resposta. No entanto, é importante reconhecer que a qualidade da invariância é uma propriedade do modelo e não dos dados. Em qualquer instância específica, precisa-se verificar se os dados se conformam com o modelo, o que justifica seu uso. No exemplo apresentado, a razão de $\hat{b}_5^* / \hat{b}_1^*$ = 9,652 foi estimada sem examinar a posição de habilidade θi* de qualquer pessoa. Se os dados são consistentes com o modelo, esta razão deve permanecer invariante ao longo da escala dos escores totais rk = 1 .. 6. Neste particular, os modelos de resposta ao item vão de encontro com a preocupação de Guttman ao examinar o ajuste do modelo.

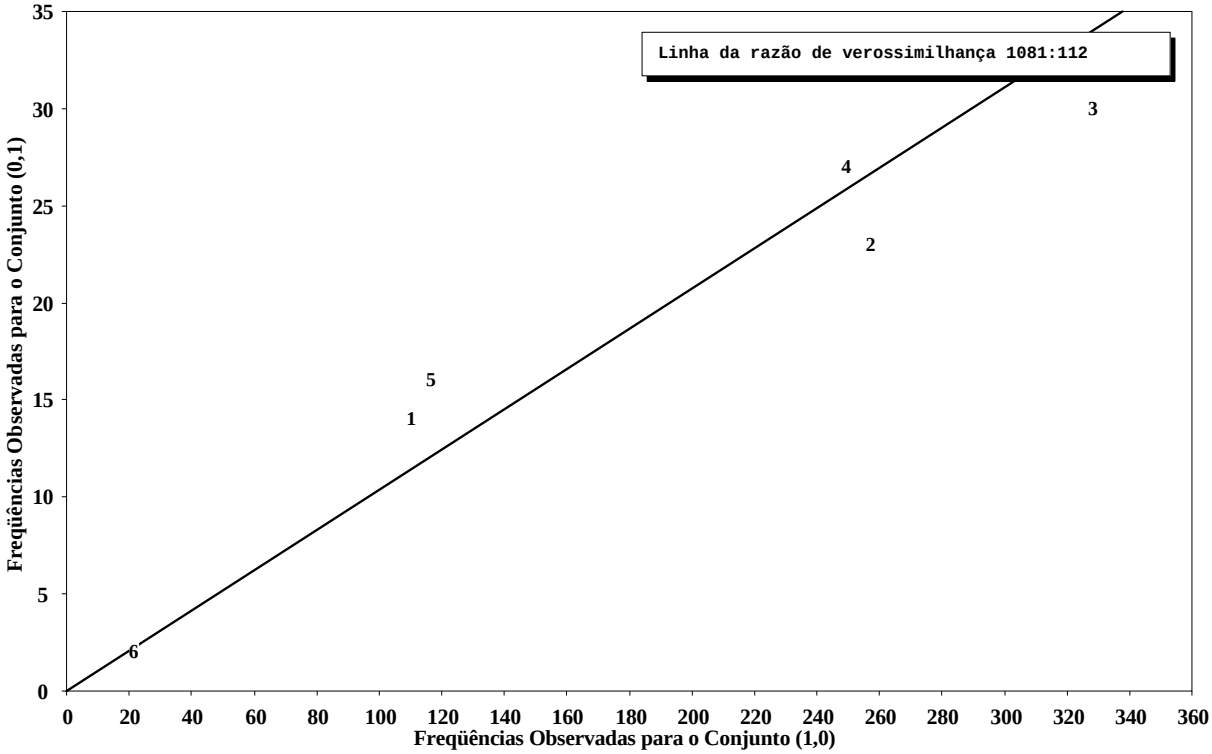
Examinamos essa questão a partir de uma comparação das freqüências observadas e esperadas apresentadas na Tabela III e a tendência linear da razão de verossimilhança apresentada no Gráfico 6. Conforme se vê na tabela, em cada nível de habilidade, as freqüências observadas dos dois itens aproximam-se razoavelmente bem dos valores

esperados baseados na razão global de $\hat{b}_5^* / \hat{b}_1^*$ = 1081 / 112 = 9,652. Para verificar se esses valores são equivalentes do ponto de vista estatístico, o teste convencional de $\chi 2$ para tabelas de contingência pode ser aplicado. Nesse caso específico, calcula-se $\chi 2$ = 2,631, com cinco graus de liberdade. Valores dessa ordem de magnitude ou maior seriam esperados em p = 76% das vezes quando os dados de fato se conformem com o modelo. Quando essa probabilidade se torna pequena, digamos, abaixo dos 5%, existiriam motivos para suspeitar que os dados não se conformem mais com o modelo. No Gráfico 6, a inspeção visual dos dados observados confirma a relação linear prevista no modelo. No caso dos dois itens de prova examinados em nosso exemplo, nada sugere que os dados sejam inconsistentes com o modelo de resposta Rasch.

Gráfico 6

Freqüência de Acerto dos Itens 1 e 5 nos Escores r_k = 1 .. 6:

Comparações Invariantes em Todos os Níveis de Habilidade



A invariância demonstra que a calibração dos itens é independente da amostra e as medidas de habilidade latente independentes dos itens. Uma prova com estas características se assemelha a um termômetro que mede a temperatura de uma substância. Os intervalos da escala do termômetro não variam de acordo com a temperatura da substância e a temperatura registrada para qualquer substância não depende do termômetro em particular que foi aplicado na ocasião. Por contraste, as provas elaboradas de acordo com o modelo clássico não dispõem de qualquer dessas características. O resultado de prova obtido por uma pessoa não pode ser interpretado sem antes fazer referência a algum grupo tido como norma e alguma forma de prova em particular aplicada na ocasião. Dessa maneira, os modelos de características latentes alcançam uma objetividade de medidas sem paralelo na teoria clássica.

A simplicidade do modelo Rasch de um parâmetro empresta uma grande elegância à análise dos dados das provas. Em determinadas circunstâncias, quando a acerto casual é

reduzida ao mínimo, os dados podem se coadunar bem com o modelo. Inclusive pode ser o único modelo aplicável em situações onde o número de provas é muito limitado. Na abordagem desse modelo, aproveitamos da razão de verossimilhança e das comparações entre pares de itens e de pessoas para exemplificar os princípios da separação e da invariância. É possível proceder com a estimação dos parâmetros dos outros itens e pessoas nessa mesma base, mas esse método não seria robusto com muitos itens, devido às pequenas freqüências de resposta encontradas em algumas das células. Por esta razão, esse método não se adapta bem à tarefa de verificar se os dados se conformam ao modelo. Normalmente se usa as probabilidades de item em vez da razão de verossimilhança e um procedimento de estimação do campo da estatística, sobretudo, a estimação segundo princípios de *máxima verossimilhança*. Este é um procedimento que procura os valores de parâmetro que fazem um conjunto de dados observados aparecer mais verossímil à luz de um modelo matemático específico. Dentro desse quadro, é possível identificar duas grandes alternativas, que se referem à estimação condicional e incondicional dos parâmetros de item.

Sob a perspectiva condicional, as pessoas e suas habilidades são tratadas como fixas, enquanto se desenvolvem estimativas dos parâmetros de item condicionais num conjunto de escores de habilidade em particular. Pressupõe-se que as respostas de diferentes itens são condicionalmente independentes entre pessoas da mesma habilidade θ. Tendo em vista que a probabilidade conjunta de eventos independentes é o produto das probabilidades dos eventos em separado, essa pressuposição permite o cálculo da probabilidade do vetor de escores $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, nas respostas de uma pessoa com habilidade θ. Partindo da probabilidade da ocorrência de uma única resposta de item em (20), pressupondo a independência entre respostas, pode-se expressar a probabilidade do conjunto x nos seguintes termos:

$$L(\boldsymbol{\beta}) = \Pr\{\mathbf{x} \mid \theta_i\} = \prod_{j=1}^{n} P_j(\theta_i)^{x_{ij}} Q_j(\theta_i)^{1-x_{ij}}, \qquad (24)$$

ou seja, o produto contínuo de $P_j(\theta_i)$ ou $Q_j(\theta_i) = 1 - P_j(\theta_i)$, de acordo com o acerto ou erro da pessoa em sua resposta ao item *j*. Quando se trata de uma afirmação teórica, a quantidade em (24) é a *probabilidade* da ocorrência do conjunto **x**, condicional em θ_i, ou seja, $\Pr\{\mathbf{x} \mid \theta_i\}$. Quando se calcula o mesmo valor usando os escores de item observados, trata-se de uma função de verossimilhança dos parâmetros de habilidade θ, ou seja, $L(\boldsymbol{\beta})$. A solução de máxima verossimilhança seria simplesmente a seleção do parâmetro que maximiza o valor calculado em (24), sempre usando os dados observados.

Procedimentos de máxima verossimilhança para a estimação dos parâmetros de item e de pessoa segundo uma perspectiva condicional podem ser encontrados em Wright & Panchapakesan (1969) e Wright & Douglas (1977), Andersen (1972) e Andrich (1988). Os dados básicos usados na estimação se referem à configuração de respostas de item dos examinados, codificadas zero ou um, em todos os n itens da prova. A tarefa de estimar simultaneamente todos os parâmetros de item é idêntica à estimação do parâmetro de cada item separadamente, em seqüência. Os procedimentos usados envolvem processos iterativos que refinam as estimativas preliminares de parâmetros em sucessivas etapas, prosseguindo assim até que se convergem em 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 ou um.

O procedimento condicional pressupõe, no entanto, que o valor do examinado na escala da característica latente seja um dado fixo. No entanto, em psicometria, a característica latente é uma construção teórica e não uma realidade física encontrada logo a mão. Surge, então, a dúvida sobre o que deveria ser usado para estimar o valor da posição de escala das pessoas na primeira etapa de estimação. Uma das soluções propostas é de

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 pessoa, prossegue-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 estima-se novas estimativas de habilidade, maximizando a função de verossimilhança. Repete-se esse ciclo iterativo, até que as estimativas dos valores de parâmetro de item se estabilizam, permanecendo essencialmente iguais de uma etapa a outra. O procedimento é curioso porque a variável considerada dependente nessa relação não é fixa e muda em função das variáveis independentes a cada etapa de estimação. No entanto, é possível obter resultados ainda razoáveis com este método. Um programa desse tipo é o LOGIST de Wingersky, Barton & Lord (1982), baseado em princípios de máxima verossimilhança conjunta propostos inicialmente por Birnbaum (1968).

O procedimento recebe críticas na literatura devido a um viés que aparece devido à estimação conjunta dos parâmetros de item e de habilidade. O viés dos estimadores não desaparece quando a amostra aumenta sem limite porque cada nova pessoa observada requer uma nova estimativa de habilidade. Os parâmetros obtidos seriam estimativas inconsistentes, quer dizer, não se aproximam de seus valores verdadeiros em amostras muito grandes. Felizmente, no caso do modelo Rasch, o escore total rk constitui uma estatística suficiente para a estimação da habilidade. Em vez de calcular $P_j(\mathbf{x}\,|\,\theta_i)$, é possível substituir $P_j(\mathbf{x}\,|\,r_k)$ e chegar a resultados consistentes que se aproximam dos valores reais quando o tamanho da amostra aumenta (Andersen, 1972). Estimativas consistentes para os outros modelos de resposta ao item dependem dos métodos incondicionais, especialmente o método de máxima verossimilhança marginal proposto por Bock & Aitkin (1981), assim como implementado subseqüentemente no programa de computador BILOG (Mislevy & Bock, 1983).

3. Referências bibliográficas

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. Journal of the Royal Statistical Society, Series B, 34, 42-54.

Andrich, D. (1988). Rasch Models for Measurement. Newbury Park, CA: Sage Publications.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. Psychometrika, 46, 443-459.

Hambleton, R. K. (1989; 1993). Principles and Selected Applications of Item Response Theory. In Linn, R. L., ed., Educational Measurement, 3rd ed. New York: American Council on Eduction. Phoenix, AZ: Oryx Press.

Hambleton, R. K., Swaminathan, H., & Rogers, H.J. Fundamentals of Item Response Theory. Newbury Park, CA: Sage Publications, 1991.

Linden, W. J. van der & Hambleton, R. K., eds (1997). Handbook of Item Response Theory. New York: Springer-Verlag.

Lord, F. M. (1980). Applications of item response theory to practical testing problems.

Hillsdale, NJ: Lawrence Erlbaum.

Mislevy, R., & Bock, R. D. (1982). BILOG: Maximum likelihood item analysis and test

scoring with logistic models. Mooresville, IN: Scientific Software.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm.

Applied Psychological Measurement, v.16, n.2 (June), 159-176.

Rasch, G. (1960; 1980). Probabilistic models for some intelligence and attainment tests.

Copenhagen: Danish Institute for Educational Research. Chicago: MESA Press.

Wright, B. D., & Douglas, G. A. (1977). Conditional versus unconditional procedures for sample-free analysis. Educational and Psychological Measurement, 37, 573-586.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. Educational and Psychological Measurement, 29, 23-48.

Wright, B. D., & Stone, M. H. (1979). Best test design. Chicago: MESA Press.