

Avaliação Multidimensional de Desempenho do Estudante*

CLAUDETTE MARIA MEDEIROS VENDRAMINI¹

Recebido em 13/07/05

Aprovado em 08/08/05

¹ Estatística, Doutora em Educação pela Universidade Estadual de Campinas e docente da graduação e do Programa de Pós-Graduação em Psicologia da Universidade São Francisco.

Endereço para correspondência: Rua Herculano Pupo Nogueira, 309, Vila Belém, Itatiba-SP, CEP 13256-300. E-mail: cvendramini@uol.com.br; claudette.vendramini@saofrancisco.edu.br

Resumo: O objetivo deste texto é tecer considerações sobre o Exame Nacional de Avaliação do Desempenho dos Estudantes (*ENADE*), e propor um modelo multidimensional de análise de mudanças de aprendizagem. São feitas considerações sobre o processo de avaliação de desempenho estudantil, *ENADE-2004*, e sobre as limitações de análise decorrente do método estatístico utilizado para avaliar os resultados obtidos por amostragem. No decorrer do texto são apresentados três modelos matemáticos da Teoria de Resposta ao Item (*TRI*) e um modelo multidimensional como proposta de análise de desempenho dos estudantes, abordando limitações do modelo proposto inicialmente e vantagens de se utilizar o novo modelo. A descrição do Modelo de Rasch Multidimensional para Aprendizagem e Mudança (*MRMLC*) foi feita com base no texto de Embretson (1996).

Palavras-chave: ENADE; modelo multidimensional de análise; teoria de resposta ao item; desempenho de estudante.

Abstract: The objective of this text is to discuss Enade, the Brazilian National Exam of Student Performance, and to propose a multidimensional model of analysis of changes in learning. Considerations are made on the student performance evaluation process, Enade – 2004, and on the limitations of the analysis due to the statistical method used to evaluate the results obtained through sampling. Throughout the text three mathematical models of the Item Response Theory and a multidimensional model as a proposal for analysis of student performance are presented, dealing with limitations of the initially proposed model and with the advantages of using the new model. The description of the Rasch Multidimensional Model for Learning and Change (*MRMLC*) was made using Embretson's text (1996) as a basis.

Key words: ENADE; Multidimensional analysis model; Item response theory; Student performance.

Introdução

O Ministério de Educação e Cultura (*MEC*) desde a década de 80 vem acumulando esforços no sentido de avaliar a educação superior no Brasil. Na década de 90 foram realizadas as primeiras experiências de avaliação em instituições públicas com

* Texto apresentado em 24 de novembro de 2004 no Encontro da Comissão do Exame Nacional de Desempenho dos Estudantes (*ENADE*), realizada no Instituto Nacional de Pesquisas Educacionais (*INEP*) do Ministério de Educação e Cultura (*MEC*), Brasília.

perspectiva formativa, como o Programa de Avaliação Institucional das Universidades Brasileiras (*PAIUB*), cuja preocupação era o processo e a missão das instituições na sociedade. Nessa mesma década foram progressivamente implementados mecanismos de avaliação que funcionaram como instrumento de classificação das instituições de ensino superior incentivando-as à concorrência entre elas. Entre esses mecanismos foi implementado o Exame Nacional de Cursos (*ENC*), do qual participaram os concluintes dos cursos de graduação. O *ENC* tinha como foco a dimensão do ensino dos cursos de graduação, e classificava os cursos com base na lógica segundo a qual a qualidade de um curso é igual à qualidade de seus alunos (*INEP*, 2004).

O *ENC*, também conhecido como provão, tinha como objetivo orientar as ações do *MEC* no sentido de estimular e fomentar a melhoria da qualidade de ensino e apoiar deliberações do Conselho Nacional de Ensino (*CNE*) quanto ao reconhecimento e renovação de cursos de graduação. Esse exame era composto por uma prova de conhecimento (provas) constituído de questões de múltipla escolha e questões discursivas, ou de questões exclusivamente discursivas, conforme adoção de cada comissão de curso. Os resultados das primeiras versões do *ENC* eram interpretados segundo a ordenação dos desempenhos, média geral dos graduandos do curso, em cinco grupos: A - 12% dos cursos com desempenho mais altos; B - 18% dos cursos com desempenho alto-médio; C - 40% dos cursos com desempenho médio; D - 18% dos cursos com desempenho médio-baixo; e E - 12% dos cursos com desempenho mais baixo. A partir de 2001 os conceitos foram baseados na conversão dos desempenhos segundo a média geral e desvio-padrão de cada área avaliada, sendo atribuído o conceito: A - cursos com desempenho igual a um ou mais desvios padrões acima da média geral; B - cursos com desempenho entre meio (inclusive) e um desvio padrão da média geral; C - cursos com desempenho entre meio desvio padrão abaixo e meio desvio-padrão acima (inclusive) da média geral; D - cursos com desempenho entre um e meio (inclusive) desvios padrões abaixo da média geral; E - cursos com desempenho abaixo ou igual a um desvio padrão da média geral (*INEP*, 2004).

Utilizado para verificar aquisição de conteúdos e medir competências dos estudantes, o *ENC* indicava a qualidade dos cursos a partir do desempenho dos graduandos nos exames, sem, no entanto, se constituir de uma medida de avaliação da aprendizagem, quanto muito podiam ser considerados como desempenho de estudantes em uma dada circunstância. A partir dos resultados do exame não era possível inferir nada sobre o valor que as instituições de ensino agregavam aos estudantes e aos cursos, valor que poderia contribuir muito para a avaliação educacional.

Além disso, o tipo de tratamento estatístico utilizado limitava as possibilidades de interpretação dos resultados e permitia que conclusões errôneas fossem tiradas sobre a qualidade dos alunos e das instituições de ensino superior. Algumas dessas limitações ainda permanecem no atual sistema de avaliação da educação superior que serão comentadas a seguir.

O Exame Nacional de Desempenho dos Estudantes - ENADE - 2004

Diante da necessidade de responder às questões da sociedade em geral quanto à validade das inferências feitas sobre a qualidade dos graduandos e dos cursos avaliados que contribuía para o reconhecimento ou renovação dos cursos foi elaborada uma nova proposta para a avaliação da educação superior. Uma Comissão Especial de Avaliação (CEA) buscou na nova proposta “consolidar as necessárias convergências em relação a uma concepção de avaliação como um processo que efetivamente vincule a dimensão formativa a um projeto de sociedade comprometido com a igualdade e a justiça social” (INEP, 2004, p.83).

A proposta de um Sistema Nacional de Avaliação do Ensino Superior (SINAES) objetiva principalmente a melhoria da qualidade da educação superior brasileira, promovendo uma avaliação das instituições de educação superior, dos cursos de graduação e do desempenho acadêmico de seus estudantes. Nesse sistema a avaliação do desempenho dos estudantes tem como um de seus objetivos acompanhar o processo de aprendizagem e o desempenho dos estudantes em relação aos conteúdos programáticos que compõem as grades curriculares dos cursos.

O Exame Nacional de Avaliação do Desempenho dos Estudantes (ENADE), que compõe esse sistema, foi realizado pela primeira vez no dia 7 de Novembro de 2004 e envolveu 14 áreas: Agronomia, Educação Física (Licenciatura e Bacharelado), Enfermagem, Farmácia, Fisioterapia, Fonoaudiologia, Medicina, Medicina Veterinária, Nutrição, Odontologia, Serviço Social, Terapia Ocupacional e Zootecnia. Os estudantes que participaram do exame foram selecionados por meio de procedimentos amostrais entre alunos ingressantes e concluintes dos cursos de graduação.

De acordo com dados da Diretoria de Estatísticas e Avaliação da Educação Superior (DEAES) do INEP foram inscritos cerca de 95.765 ingressantes e 59.888 concluintes de 2.184 cursos de graduação. O curso com maior número de participantes inscritos foi Educação Física, 32.682 alunos, e o curso com menor número foi Terapia Ocupacional, com 1.963 estudantes. Esses alunos foram selecionados aleatoriamente a partir de uma relação completa de estudantes ingressantes e concluintes enviadas pelas instituições de ensino.

A realização do exame por alunos selecionados aleatoriamente por amostragem estratificada e a obrigatoriedade na inscrição de alunos iniciantes e concluintes parece resolver um dos problemas do exame anterior, o boicote dos alunos ao exame. O processo de amostragem aleatória e a representativa da amostra permitem que sejam feitas inferências confiáveis de características educacionais da população, que possam orientar a tomada de decisões relacionadas aos cursos avaliados.

Neste novo sistema de avaliação as provas são compostas por 10 questões de avaliação da formação geral, oito de múltipla escolha e duas discursivas, comuns aos cursos de todas as áreas, e 30 questões específicas para cada área, também compostas por questões discursivas e de múltipla escolha. Tanto alunos ingressantes quanto concluintes respondem as mesmas 40 questões formuladas para cada curso especificamente.

O aperfeiçoamento da nova proposta de exame pode permitir inferências confiáveis sobre o valor que as instituições de ensino agregam aos estudantes e que não era possível de ser feita no sistema de avaliação anterior. Esse aperfeiçoamento inclui avaliações sucessivas do processo de aprendizagem, que levem em consideração as habilidades iniciais dos estudantes como também as habilidades efetivas desenvolvidas por eles durante o curso, portanto uma avaliação multidimensional de mudanças educacionais. Assim, uma parcela de ingressantes deveria ser avaliada no início, durante e no final do processo de aprendizagem, permitindo a extração do valor agregado como indicador de ganho, um dos objetivos do novo exame, o *ENC*. Mesmo avaliando apenas o início e o final de curso, esse último considerado como momento final da aprendizagem, essas medidas de mudança poderão orientar tomadas de decisões públicas e institucionais.

Medidas de mudanças educacionais, como as referentes ao valor agregado, pressupõem segundo a literatura educacional, que se avalie o conhecimento prévio a respeito do estudante (Ferrão, 2003; Sternberg & Grigorenko, 2002; Embretson, 1996). Espera-se que no final do curso o nível de habilidade efetiva dos estudantes, avaliada pelo desempenho observado em provas ou exames válidos para esse fim, atinja níveis mais elevados que no início do curso, indicando ganho para os alunos, curso e instituição. Melhores resultados em provas equivalentes às realizadas pelos estudantes no início do curso refletem o que o indivíduo é capaz de fazer com o conhecimento aprendido na escola e não apenas “o que” e “quanto” ele aprendeu, que somado a outros indicadores obtidos com a aplicação dos vários instrumentos e ações que compõem o *SINAES*, permitirão identificar quais aspectos da *IES* e dos cursos precisam ser melhorados ou mantidos.

As medidas de avaliação do valor agregado e desempenho estudantil - *ENC* - 2004

A avaliação de resultados obtidos em provas por meio de escores brutos ou padronizados, como proposto pelo *ENC*, apresenta algumas limitações, como por exemplo, ser dependente do conjunto particular de itens que compõem a prova. As análises e interpretações estão associadas ao resultado geral da prova (pontuação total), característica principal da teoria clássica de medida. Embora, os ingressantes e concluintes realizem a mesma prova, tornando possível comparar os resultados desses indivíduos, essa teoria não permite avaliar o valor agregado ao estudante, por não ser possível inferir sobre a sua habilidade inicial, ou seja, habilidade efetiva no início do curso de graduação.

Mesmo sob a suposição de que os concluintes de hoje teriam em seu ingresso na universidade uma habilidade inicial correspondente aos ingressantes de hoje, a proposta de medida do valor agregado, calculado a partir da diferença entre as médias, não permite fazer inferências sobre o valor agregado, isto é, sobre as mudanças ocorridas no indivíduo, dado que não se considera na análise o padrão de respostas aos itens da prova de cada um dos indivíduos.

Diante do exposto propõe-se fazer uso da Teoria de Resposta ao Item Multidimensional para avaliar as mudanças ocorridas no processo educacional, conforme apresentado a seguir.

A Teoria de Resposta ao Item (*TRI*)

A *TRI*, é um procedimento de medida utilizado sob a suposição de que existe no indivíduo um traço (uma característica individual determinante de como responder aos itens de uma prova) que possui uma relação probabilística com cada um dos itens utilizados. Considerando-se que os parâmetros de cada item não dependem dos outros itens da prova, mas que a pontuação da prova se faz em função das respostas do indivíduo a cada item, é possível verificar se os respondentes são mais ou menos hábeis, e da mesma forma, se os itens podem ser considerados mais fáceis ou mais difíceis, já que itens e indivíduos são colocados na mesma escala de desempenho.

A *TRI* não entra em contradição com os princípios da psicometria clássica e traz uma nova proposta estatística, a de análise centrada nos itens, que supera as limitações da teoria clássica, além de apresentar novos recursos tecnológicos para a avaliação psicológica e educacional (Primi, 1998). Existem vários modelos possíveis de resposta ao item, que diferem em sua forma em função da característica do item e do número de parâmetros especificados no modelo. Todos os modelos dessa teoria possuem um ou mais parâmetros que descrevem o item e um ou mais parâmetros que descrevem o respondente. O primeiro passo para uma aplicação da *TRI* é a estimação desses parâmetros (Hambleton, Swaminathan & Rogers, 1991).

Alguns pontos têm sido levantados na literatura sobre a adequação desta teoria na área de avaliação educacional, dois deles, considerados importantes são: a dimensionalidade do espaço de traços latentes envolvidos na avaliação e a equalização de diferentes avaliações. É necessário ressaltar que, apesar de não haver dúvidas de que a aplicação desta teoria muito contribui para a melhoria das avaliações educacionais em geral, sua disseminação efetiva depende da integração de especialistas das áreas de estatística e educação (Andrade, 2001).

Uma das vantagens da utilização da *TRI* na avaliação educacional é que esta possibilita análises qualitativas a partir dos resultados brutos de uma prova, fornecendo assim informações mais precisas do desempenho dos respondentes e da qualidade das questões utilizadas (itens), questões que devem ter índices de dificuldade e de discriminação aceitáveis e correlacionadas com a prova total. No Brasil, a *TRI* vem sendo aplicada em diversas avaliações educacionais desde 1995, como na análise de dados do Sistema Nacional do Ensino Básico - *SAEB* e do Sistema de Avaliação do Rendimento Escolar do Estado de São Paulo - *SARESP* (Vendramini, 2002). Nessas avaliações, o objetivo é comparar o desempenho dos alunos em diferentes séries e disciplinas, utilizando questões comuns entre ciclos de aplicação e entre séries.

Convém, no entanto, ressaltar que os modelos de resposta ao item só podem ser considerados vantajosos quando o ajuste do modelo aos dados de interesse for satisfatório. Um modelo mal-ajustado não fornecerá parâmetros invariantes para os itens e para as habilidades (Hambleton, Swaminathan & Rogers, 1991).

Tanto para a análise pela Teoria Clássica dos Testes (*TCT*) quanto pela *TRI* são indicadas as respostas dos estudantes nas cinco alternativas de cada item, e as respostas corretas, para possibilitar que os itens assumam escores do tipo certo/errado (itens dicotômicos). Embora as alternativas dos itens da prova do *ENADE* não tenham sido construídos com o objetivo de evidenciar raciocínios falhos dos estudantes, a análise dos índices de discriminação e das correlações ponto bisserial (item-total) podem revelar tendências de escolha da opção errada em alguns dos itens, fornecendo indicadores de raciocínio errado dos estudantes, o que pode auxiliar no processo de avaliação da aprendizagem.

Principais modelos matemático

Os modelos matemáticos propostos na literatura variam conforme a natureza da questão ou item (dicotômico ou não dicotômico), o número de populações envolvidas (uma ou mais de uma) e o número de traços latentes (habilidades) que estão sendo medidos (um ou mais de um).

Antes de empregar os modelos matemáticos da *TRI*, deve-se comprovar o cumprimento de dois pressupostos teóricos fundamentais. Para a utilização de modelos unidimensionais, os critérios são: (1) *critério da unidimensionalidade*: os itens de uma prova devem medir uma única habilidade, ou ao menos deve haver um fator dominante que influencie o desempenho dos respondentes na prova; (2) *critério da independência local*: as respostas dos participantes aos itens não devem ser influenciadas pelas respostas fornecidas a outros itens, para indivíduos com uma mesma habilidade.

Os parâmetros dos itens mais relevantes são: a *dificuldade*, a *discriminação* e a *probabilidade de acerto por acaso* (isto é, a probabilidade de um indivíduo de baixa habilidade dar uma resposta correta a um item difícil). A *dificuldade* do item é dada na mesma escala da habilidade, e refere-se à habilidade necessária para uma dada probabilidade de acertar o item, calculada a partir da probabilidade de acertar o item por acaso. A *discriminação* corresponde à inclinação da Curva Característica do Item (*CCI*) e indica quanto indivíduos de diferentes habilidades diferem quanto à probabilidade de acertar um item. A representação gráfica da *CCI* tem forma de “S” (Figura 1), com inclinação e deslocamento na escala de habilidade definidos pelos parâmetros dos itens. No eixo das abscissas está indicado o nível observado de habilidade (traço latente) do indivíduo, designado por q , cujo valor pode variar de $-\infty$ a $+\infty$, e no eixo das ordenadas a probabilidade de responder corretamente ao item i , designado por $P_i(q)$ e variando de 0 a 1. Assim, os modelos matemáticos da *TRI*, representam para cada item i a função de probabilidade de indivíduos com habilidade q acertarem esse item (Vendramini, Silva, & Canale, 2004).

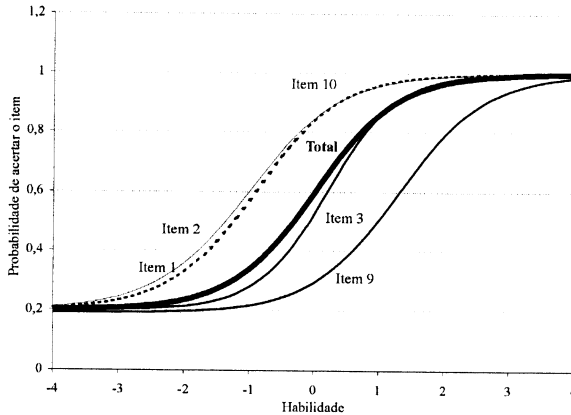


Figura 1 - Curvas características de alguns itens de uma prova (Vendramini, Silva & Canale, 2004)

Na TRI, a probabilidade de resposta correta depende da habilidade do indivíduo j (θ_j), que permite expressar a sua resposta aos itens, e dos parâmetros do item i (a_i , b_i e c_i), sendo, portanto, necessário estimar valores destes parâmetros que melhor expliquem os resultados obtidos, com base nas respostas dos indivíduos aos itens. Esse processo é chamado de calibração ou parametrização, e é feito com o auxílio de programas específicos (*RASCAL*, *XCALIBRE*, *BILOG*, entre outros).

Nas expressões dos modelos matemáticos apresentados a seguir, a probabilidade condicional $P(X_{ij}=1|\theta_j)$ é denotada apenas por $P_i(\theta)$. Para cada item i a probabilidade de acerto é uma função da habilidade θ do indivíduo, que varia de $-\infty$ a $+\infty$. Conhecidos a habilidade θ de um indivíduo j e os parâmetros do item i é possível determinar a probabilidade de ele acertar esse item pelas expressões (1), (2) ou (3) de acordo com o modelo escolhido.

Modelo logístico de três parâmetros

O modelo de três parâmetros é o modelo teórico mais completo. Considera como variáveis que influenciam a probabilidade do indivíduo acertar o item os três parâmetros citados: a dificuldade, a discriminação e a probabilidade de acerto ao acaso, e é expresso pela função matemática (1) a seguir.

$$P(X_{ij}=1|\theta_j) = P_i(\theta) = c_i + \frac{(1-c_i)}{1 + e^{-Da_i(\theta-b_i)}} \quad (1)$$

com $i = 1, 2, \dots, n$ e $j = 1, 2, \dots, m$, sendo:

X_{ij} uma variável dicotômica que assume os valores 1 ou 0, conforme o indivíduo j responda correta ou incorretamente o item i , respectivamente;

θ_j o valor que representa a variável latente (aptidão ou habilidade) que permite explicar a resposta do j -ésimo indivíduo aos itens;

$P(X_{ij} = 1 | \theta_j) = P_i(\theta)$ a probabilidade de um indivíduo j com habilidade θ_j responder corretamente o item i ;

c_i a probabilidade de acerto ao acaso do item i ;

b_i o índice de dificuldade (ou parâmetro de posição) do item i , medido na mesma escala da habilidade θ . Corresponde à habilidade necessária para uma probabilidade de acerto igual a $(1 + c_i) / 2$;

a_i o índice de discriminação (ou parâmetro de inclinação) do item i , com valor proporcional à inclinação da CCI no ponto b_i ;

D um fator de escala constante, igual a 1 ou a 1,7 (quando se deseja que a função logística se aproxime da ogiva normal);

e um número transcendental, base dos logaritmos neperianos, cujo valor é aproximadamente 2,718;

n o número de itens;

m o número de indivíduos.

Modelo logístico de dois parâmetros

O modelo de dois parâmetros possui em sua expressão o índice de dificuldade b_i e o de discriminação a_i , como variáveis que influenciam a probabilidade de o indivíduo acertar o item. Esse modelo pode ser entendido como um modelo de três parâmetros com o valor $c_i = 0$ (Expressão 2).

$$P(X_{ij} = 1 | \theta_j) = P_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}} \quad (2)$$

Modelo logístico de um parâmetro

O modelo de um parâmetro, também conhecido como modelo de Rasch, possui em sua expressão o índice de dificuldade b_i , que se relaciona com a probabilidade de acertar o item i por acaso. Esse modelo pode ser entendido como um modelo de três parâmetros com o valor $c_i = 0$ e mesmo valor a para todos os a_i 's (Expressão 3).

$$P(X_{ij} = 1 | \theta_j) = P_i(\theta) = \frac{1}{1 + e^{-Da(\theta - b_i)}} \quad (3)$$

Análise da dimensionalidade da prova

Um das suposições da TRI é que a prova seja unidimensional, ou pelo menos que se possa assumir um fator predominante, quando se desejar fazer uso de modelos unidimensionais. O programa *TESTFACT* efetua a análise considerando questões do tipo certo/errado (dados dicotômicos) a partir das respostas dos participantes (em vez da matriz de correlação). Esta análise é denominada Análise Fatorial com Informação Completa (*Full Information Factor Analysis*) e inclui progressivamente fatores

que indicam a contribuição do fator incluído para a explicação das correlações entre os itens, possibilitando fazer previsões das respostas dos participantes aos itens a partir das curvas dos itens. Diferentes padrões de resposta podem ser esperados quando os itens são completamente independentes, ou quando medem um único fator ou mais de um fator. A adequação de um modelo uni ou multidimensional aos padrões de respostas dos participantes é verificada pelo teste Qui-quadrado. É imprescindível verificar, por este mesmo teste, se a inclusão sucessiva de fatores nos modelos tem um efeito significativo (Bock, Gibbons & Muraki, 1988).

Os resultados da análise fatorial com informação completa indicam a correlação média tetracórica entre os pares de combinações, dois a dois, dos itens. Inicialmente é verificada a variância explicada de um modelo unidimensional. Incluindo-se um segundo fator, modelo bidimensional, é verificado se a magnitude da diferença entre os padrões de resposta observados e os reproduzidos pelo modelo é estatisticamente significativa. Ao se considerar um modelo tridimensional, é verificado se a magnitude da diferença entre os padrões de resposta observados e os reproduzidos pelos modelos, quando se acrescenta um terceiro fator é significativa. E assim sucessivamente até que o acréscimo de um novo fator não seja significativo.

A seguir, é apresentado um modelo multidimensional apresentado por Embretson (1996) adequado para medir mudanças educacionais (ou o valor agregado) conforme proposta atual do *ENADE*, o modelo multidimensional de traços latentes para medir aprendizagem e mudança. Para atender à proposta do *ENADE* seriam consideradas duas habilidades: habilidade efetiva inicial e habilidade efetiva final, considerando a mudança entre esses dois momentos da aprendizagem (início e final do curso), conforme descrito a seguir.

Modelo Multidimensional de Traços Latentes para medir Aprendizagem e Mudança (*Multidimensional Rasch Model for measuring Learning and Change - MRMLC*).

Embretson (1996) descreve o modelo *MRMLC* por meio do qual a habilidade assumida para produzir um desempenho observado, denominada habilidade latente efetiva de um indivíduo (habilidade antes da mudança ser observada), é a soma de várias sub-habilidades, as habilidades de aprendizagem remanescentes. Quando o modelo se ajusta propriamente aos dados, é possível comparar itens e escala, tornando possível construir instrumentos comparáveis contendo esquemas diferentes de medida ou para uso em diferentes ocasiões.

Para Embretson (1996), é difícil desenvolver modelos psicométricos apropriados para medir processos de aprendizagem. A autora cita três problemas observados por Bereiter (1963) ao medir mudanças por diferença de escores: (a) a precisão da mudança é inversamente relacionada à precisão das provas; (b) mudanças não podem ser medidas em uma mesma escala para indivíduos de diferentes níveis de escores

iniciais; e (c) escores de mudança têm uma falsa relação negativa com os escores iniciais. Essas dificuldades parecem não ter solução quando as medidas de mudança são conceituadas dentro da *TCT*. Tanto o nível médio de mudança quanto o significado da mudança (avaliado por correlações) dependem da distribuição de escores iniciais, que é uma população específica. Resultados de pesquisas como a de Woodrow's (citado por Embretson, 1996) revelam que mudanças no desempenho estão altamente correlacionadas com o nível de desempenho inicial dos indivíduos. Esses problemas podem ser solucionados quando medidas de mudança são conceituadas dentro da *TRI*.

As medidas de habilidade do modelo *MRMLC* podem corresponder a ocasiões múltiplas ou a condições múltiplas presentes em uma única ocasião. Dois tipos de habilidades são medidas por esse modelo: *habilidade efetiva*, o potencial latente de desempenho de um indivíduo em uma ocasião particular ou em condições específicas de medida (habilidade efetiva em ocasiões iniciais é tradicionalmente denominada habilidade inicial); *habilidade de aprendizagem*, o incremento na habilidade efetiva entre duas medidas sucessivas. Em uma segunda medida (ou medida posterior) a habilidade efetiva depende da habilidade inicial e de uma ou mais habilidades de aprendizagem.

Subjacente ao desempenho está a dificuldade do item, b_i , e M habilidades, θ_{jm} para cada indivíduo j , tal que θ_{j1} é a habilidade inicial e θ_{j2} até θ_{jm} são as habilidades de aprendizagem. Assume-se que um acréscimo no desempenho de um item por um indivíduo j resulta de um acréscimo na habilidade efetiva, que é devido à mudança nas habilidades de aprendizagem, mais do que à mudança na dificuldade do item. As dificuldades dos itens são constantes em muitas condições, assim como uma única dificuldade de item, b_i , se aplica ao item com alguma condição. Assume-se que itens dentro de uma mesma condição têm discriminações denotadas por a_{km} , então o desempenho em algum item administrado sob uma condição k é a *habilidade efetiva* θ_{jk}^* . As habilidades efetivas são um conjunto de equações que representam a habilidade inicial e as habilidades de aprendizagem do indivíduo, tal como se segue:

$$\begin{aligned} \theta_{j1}^* &= a_{11} \theta_{j1} \\ \theta_{j2}^* &= a_{21} \theta_{j1} + a_{22} \theta_{j2} \\ \theta_{j3}^* &= a_{31} \theta_{j1} + a_{32} \theta_{j2} + a_{33} \theta_{j3} \\ &\dots\dots\dots \\ \theta_{jk}^* &= a_{k1} \theta_{j1} + a_{k2} \theta_{j2} + \dots + a_{km} \theta_{jm} \end{aligned}$$

De acordo com as equações anteriores, habilidades são crescentemente incorporadas na habilidade efetiva em cada condição sucessiva k . Esse modelo expressa o conceito de habilidade dinâmica e não estática.

O modelo apresenta algumas restrições: em uma mesma condição k os parâmetros de discriminação dos itens para cada dimensão m são iguais para todos os itens i .

Embora o modelo de habilidade efetiva seja multidimensional para $k > 1$, θ_k^* é uma composição ponderada de habilidades que é a mesma para todos os itens naquela condição k . Esquemáticamente a relação entre habilidade efetiva, habilidade inicial e habilidade de aprendizagem, e itens está apresentada a seguir.

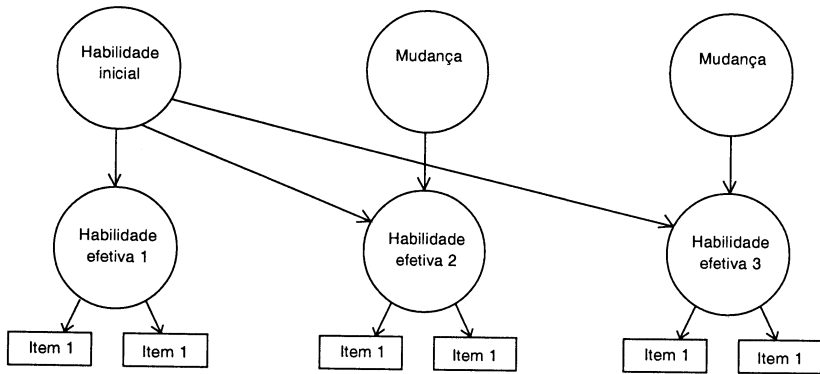


Figura 2 - Relação esquemática entre habilidade efetiva, habilidade inicial e habilidade de aprendizagem, apresentada por Embretson (1996)

Uma importante característica do modelo de Rasch, o modelo mais simples da TRI, é que as habilidades são medidas como variáveis latentes. O potencial latente x_{ij} , que um indivíduo j resolve um item i , é a diferença entre a habilidade do indivíduo θ_j^* e a dificuldade do item b_i ($\theta_j^* - b_i$).

O modelo de Rasch, modelo logístico no qual o potencial de resposta latente determina que o indivíduo resolve um item particular é dado por:

$$P(X_{ij} = 1 | \theta_j^*, b_i) = \frac{1}{1 + e^{-(\theta_j^* - b_i)}}$$

A habilidade é estimada a partir das respostas dos indivíduos aos itens da prova, tal que a diferença entre as respostas dos indivíduos e a probabilidade predita pela equação anterior é tão pequena quanto possível. A principal vantagem deste modelo sobre a TCT é que a habilidade estimada não é influenciada pela dificuldade do item.

Na TCT a diferença relativa na habilidade de dois indivíduos ou de um mesmo indivíduo em duas ocasiões, depende se os itens da prova são fáceis, moderados ou difíceis. Duas habilidades medidas pelo escore total ou por uma transformação linear (incluindo o escore z) podem parecer completamente diferentes sobre um conjunto de itens moderadamente difíceis, mas indistinguíveis sobre um conjunto de itens muito fáceis ou muito difíceis. No modelo de Rasch, a dificuldade dos itens é controlada pelas estimativas da habilidade fornecida pelo modelo. As mesmas diferenças entre duas habilidades são dadas no modelo de Rasch, para itens fáceis, moderados e

difíceis. Ao medir mudanças espera-se que as mensurações independam dos instrumentos de medida utilizados.

A *MRMLC* é uma extensão do modelo multidimensional de Rasch para dados binários, incluindo uma restrição sobre os parâmetros, isto é, que eles são constantes sobre as condições. Assume-se que as dificuldades dos itens, b_i , são constantes sobre as k condições e que a habilidade efetiva é dada como uma combinação da habilidade inicial e habilidades de aprendizagem, e que todas as discriminações dos itens não nulos, a_{km} , são unitários.

Portanto, a habilidade efetiva para uma condição k dada por:

$$\theta_{jk}^* = a_{k1} \theta_{j1} + a_{k2} \theta_{j2} + \dots + a_{km} \theta_{jm}$$

fica igual a:

$$\theta_{jk}^* = \theta_{j1} + \theta_{j2} + \dots + \theta_{jm} = \sum_{m=1}^k \theta_{jm}$$

Essa estrutura mostra que as habilidades efetivas dependem de uma soma não ponderada da habilidade inicial e das $(k - 1)$ habilidades de aprendizagem. Embora, essas habilidades possam ter um impacto diferenciado sobre o desempenho, não é necessário nenhum outro parâmetro de discriminação além da unidade.

Substituindo θ_{jk}^* em $P(X_{ij} = 1 | \theta_j^*, b_i) = \frac{1}{1 + e^{-(\theta_j^* - b_i)}}$ pela $\sum_{m=1}^k \theta_{jm}$

têm-se:

$$P(X_{ij} = 1 | \theta_j^*, b_i) = \frac{1}{1 + e^{-\left(\sum_{m=1}^k \theta_{jm} - b_i\right)}}$$

Escores de mudança em diferentes níveis da habilidade inicial

Segundo Bereiter (citado por Embretson, 1996) pequenas mudanças ocorridas quando a habilidade inicial é alta pode não ter o mesmo significado que as mesmas mudanças ocorridas quando a habilidade inicial é moderada.

Na *TCT* as mudanças de desempenho estão linearmente relacionadas aos escores de mudança estimados. Na *TRI*, a relação de habilidade para o desempenho (expresso como probabilidades de resposta) são não-lineares, e, a mudança exata na probabilidade entre duas habilidades efetivas depende de suas posições relativas na curva de resposta ao item.

No modelo *MRMLC* as mudanças esperadas no desempenho para um dado nível de habilidade de aprendizagem dependerá tanto do nível de habilidade inicial quanto da dificuldade dos itens. Por exemplo, dois indivíduos com diferentes habilidades iniciais (θ_{11} e θ_{21}) e com habilidades de aprendizagem diferentes (θ_{12} e θ_{22}) apresentam valores diferentes quanto ao acréscimo de desempenho. Na Curva Ca-

racterística do Item (*CCI*), apresentada a seguir, pode-se observar que o acréscimo de desempenho é maior para o Indivíduo 2 do que para o Indivíduo 1.

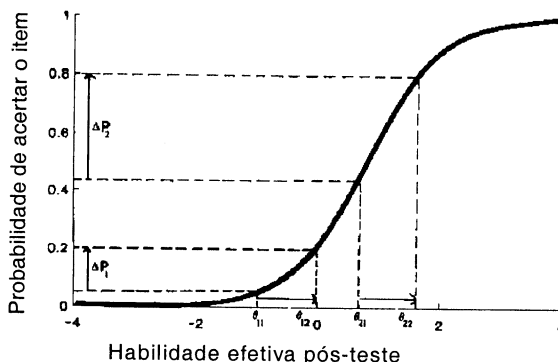


Figura 3 - Curva Característica do Item referente à habilidade efetiva no pós-teste

Em estudos realizados sobre os escores esperados de mudanças do *MRMLC* para habilidade de aprendizagem de três níveis diferentes de habilidade inicial, observou-se que níveis mais altos de habilidades de aprendizagem estão associados com mais mudanças de desempenho para níveis iniciais de habilidades mais baixos do que aos mais altos (Embretson, 1996).

Considerações finais

A avaliação da educação brasileira tem sido uma preocupação constante do Ministério de Educação e Cultura desde a década de 80. A partir da década de 90 vários mecanismos de avaliação foram progressivamente implementados, que funcionaram como instrumento de classificação das instituições de ensino superior incentivando-as à concorrência entre elas. Entre esses mecanismos foi implementado o Exame Nacional de Cursos (*ENC*), do qual participaram concluintes dos cursos de graduação, e que objetivava verificar aquisição de conteúdos ou medir competências dos estudantes, indicando a qualidade dos cursos a partir do desempenho dos graduandos nos exames, sem, no entanto, se constituir de uma medida de avaliação da aprendizagem. A partir dos resultados desse exame não era possível inferir nada sobre o valor que as instituições de ensino agregavam aos estudantes e aos cursos, resultado que poderia contribuir muito para a avaliação educacional. Diante de críticas feitas pela sociedade em geral quanto à análise e aplicação dos resultados do *ENC*, uma Comissão Especial de Avaliação (*CEA*) propõe um novo Sistema Nacional de Avaliação do Ensino Superior (*SINAES*) buscando a melhoria da qualidade da educação superior brasileira, promovendo uma avaliação das instituições de educa-

ção superior, dos cursos de graduação e do desempenho acadêmico de seus estudantes. Essa nova proposta inclui o Exame Nacional de Avaliação do Ensino Superior cuja análise é feita por métodos da teoria clássica de medida.

Diante das limitações de análise e de inferências sobre o desempenho dos estudantes e cursos obtidos no *ENADE*, propõe-se um modelo multidimensional para avaliar aprendizagem e mudança, a partir do modelo multidimensional de Rasch, descrito por Embretson (1996). Espera-se com esta proposta contribuir para o aperfeiçoamento do *ENADE* permitindo inferências mais abrangentes sobre as mudanças educacionais ocorridas na universidade, a partir de indicadores de ganho e de desempenho de estudantes e cursos de graduação.

Referências

- Andrade, D. F. (2001). Comparando desempenhos de grupos de alunos por intermédio da teoria da resposta ao item. *Estudos em Avaliação Educacional*, 23, 31.
- Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full Information item factor analysis. *Applied Psychological measurement*, 12(3), 261-280.
- Embretson, S. E. (1996). Implications of a multidimensional latent trait model for measuring change. In Collins, Linda M. (Ed); Horn, John L. (Ed). *Best methods for the analysis of change: Recent advances, unanswered questions, future directions*. (pp. 184-197). Washington, DC, US: American Psychological Association.
- Ferrão, M. E. (2003). *Introdução aos modelos de regressão multinível em Educação*. Campinas, SP: Komedi.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park: Sage Publishers.
- Instituto Nacional de Pesquisas Educacionais Anísio Teixeira (2004). *Sistema Nacional de Avaliação da Educação Superior: da concepção à regulamentação*. 2ª ed., ampl. Brasília: INEP.
- Primi, R. (1998). *Desenvolvimento de um instrumento informatizado para a avaliação do raciocínio analítico*. Tese Doutorado. Universidade de São Paulo.
- Sternberg, R. J., & Grigorenko, E (2002). *Dinamic learning*. New York: Cambridge University Press.
- Vendramini, C. M. M. (2002). Aplicação da teoria de resposta ao item na avaliação educacional. In: Primi R. (Org.). *Temas em avaliação psicológica* (pp.116-130). Campinas: Instituto Brasileiro de Avaliação Psicológica.
- Vendramini, C. M. M., Silva, M. C., & Canale, M. (2004). Análise de itens de uma prova de raciocínio estatístico. *Psicologia em Estudo*, 9(3), 331-342,