TEORIA DA RESPOSTA AO ITEM: UMA ANÁLISE CRÍTICA DOS PRESSUPOSTOS EPISTEMOLÓGICOS

CRISTINA ZUKOWSKY TAVARES

RESUMO

O artigo procura desvelar concepções e valores envolvidos no ato de avaliar e sua interface com a ideia de medida, tomando como referência alguns estudos disponíveis na literatura educacional sobre a Teoria da Resposta ao Item, a fim de verificar suas potencialidades e limites. Investigou-se a Teoria da Resposta ao Item sob o prisma de três relevantes pressupostos epistemológicos do modelo: unidimensionalidade, independência local e qualidade e adequação dos itens. Algumas questões podem ser levantadas com base nesta análise dos postulados: "Como garantir a unidimensionalidade quando todo comportamento humano é altamente complexo e multideterminado?"; "E elaborar questões de provas totalmente independentes?"; "É possível construir e manter um banco de itens calibrados e adequados?". Esses são importantes desafios para a concretização da teoria e a confiabilidade.

PALAVRAS-CHAVE TEORIA DA RESPOSTA AO ITEM • ESCALA DE AVALIAÇÃO • EPISTEMOLOGIA • AVALIAÇÃO DA EDUCACIONAL.

RESUMEN

El artículo pretende mostrar concepciones y valores implicados en el acto de la evaluación y su interfaz con la idea de medida, tomando como referencia algunos estudios disponibles en la literatura educativa sobre la Teoría de la Respuesta al Ítem, con el fin de verificar sus potencialidades y sus límites. Se investigó la Teoría de la Respuesta al Ítem a partir de tres presupuestos epistemológicos relevantes del modelo: la unidimensionalidad, la independencia local y la calidad y adecuación de los ítems. El análisis de estos postulados permite plantear algunas preguntas: ¿Cómo garantizar la unidimensionalidad cuando todo comportamiento humano es altamente complejo y multideterminado? ¿Cómo elaborar preguntas de pruebas totalmente independientes? ¿Es posible construir y mantener una base de ítems ajustados y adecuados? Estos son importantes desafíos –que no siempre se presentan como infalibles– para la concreción de la teoría y la confiabilidad de su metodología.

PALABRAS CLAVE TEORÍA DE LA RESPUESTA AL ÍTEM • ESCALA DE EVALUACIÓN • EPISTEMOLOGÍA • EVALUACIÓN EDUCACIONAL.

ABSTRACT

This article seeks to reveal concepts and values involved in the act of evaluating and its interface with the idea of measurement. It also addresses some studies on the Item Response Theory (IRT) in the literature in order to assess its potential and limits. IRT was investigated from the perspective of three important epistemological assumptions of the model: unidimensionality, local independence, and quality and appropriateness of items. "How can unidimensionality be maintained if all human behavior is highly complex and multidimensional?"; "How can one prepare test items which are totally independent?"; "Is it possible to create and maintain a database with calibrated and adequate items?". These are important challenges for the theory and the reliability of the methodology, which are not always infallible.

KEYWORDS ITEM RESPONSE THEORY • RATING SCALE • EPISTEMOLOGY • EDUCACIONAL EVALUATION.

INTRODUÇÃO

Nas últimas décadas, a ideia de avaliação educacional tem-se situado no centro das atenções nos diferentes níveis de ensino. Desde o início dos anos 1990, sobretudo nos países anglo-saxônicos, as funções mais importantes atribuídas à avaliação educacional são, essencialmente, as que remetem à seleção dos indivíduos e à "gestão produtivista" do sistema educativo. Na perspectiva de mercado educacional, a avaliação, fundamentalmente, auxilia as escolhas dos consumidores da educação. Algumas vezes, esses resultados são analisados com cuidado, refletidos individual e coletivamente, subsidiando e apoiando ações de melhoria pelos agentes educacionais envolvidos. No entanto, o que ocorre em outros momentos é que esses dispendiosos exames e avaliações em larga escala não são compreendidos, valorizados e bem aproveitados pelos que estão diretamente envolvidos, ou, mesmo, se encontram no nível da consecução e planejamento de políticas públicas de intervenção na educação. O exame se tornou um instrumento por meio do qual os agentes educacionais – mesmo nesta época – têm a "esperança de melhorar a educação", considerando

que "existe uma relação simétrica entre sistema de exames e sistema de ensino". Acredita-se, assim, no falso princípio de que "um melhor sistema de exame" garante "melhor sistema de ensino". Tudo isso é muito falso já que o exame é "um efeito das concepções sobre aprendizagem, não o motor que transforma o ensino". Há uma excessiva confiança no exame como instrumento capaz de melhorar a qualidade da educação e de toda a vida social, como acentua Barriga:

> Porém, o exame é só um instrumento que não pode por si mesmo resolver os problemas gerados em outras instâncias sociais. Não pode ser justo quando a estrutura social é injusta; não pode melhorar a qualidade da educação quando existe uma drástica redução de subsídio e os docentes se encontram mal pagos; não pode melhorar os processos de aprendizagem dos estudantes quando não se atende nem à conformação intelectual dos docentes, nem ao estudo dos processos de aprender de cada sujeito [...] afirmamos que o exame é um espaço social superdimensionado. Também enunciamos que o exame não pode resolver uma infinidade de problemas que se condensam nele. (2003, p. 27)

AVALIAÇÃO: A IDEIA DE MEDIDA E DE VALOR

A avaliação educacional está relacionada a questionamentos na ordem dos valores e não pode se reduzir a uma medida objetiva de desempenhos. Um processo de avaliação nunca se esgota em um processo de medida, vai além dele. A despeito da inserção da medida no universo da avaliação, na expressão de Machado (1997, p. 31), a avaliação não se restringe à ideia de medida da mesma forma que

> [...] a pessoalidade inclui a cidadania como seu núcleo duro, mas é muito mais abrangente; assim como a cidadania inclui o consumidor, mas não se esgota nele; assim como o conhecimento inclui a dimensão mercantil, mas a transcende em muito.

E assim também a avaliação transcende, transborda a noção de medida. Então, até que ponto avaliar é medir?

> Como se vê, embora a ideia de avaliar inclua a de calcular, computar, determinar o preço, expressar numericamente ou

> expressar o valor matemático de algo, em seu cerne encontra-se, sem dúvida, a ideia de valor, a emissão de juízos de valor [...]. Esta importante dimensão do significado da avaliação – o caráter e a complexidade de um julgamento – resulta, no entanto, frequentemente subestimada em certas vertentes do discurso educacional e na quase totalidade do discurso político relativo ao tema, superestimando-se a caracterização da avaliação como um processo de medida de natureza essencialmente técnica. Assim, por meio do recurso ao biombo da medida, é evitada ou minimizada grande parte das dificuldades inerentes ao enfrentamento da ideia de valor, ainda que o custo de tal alívio seja um esvaziamento no significado das reflexões sobre o tema. (MACHADO, 2006, p. 72-73)

A preocupação com a objetividade predominou por alguns séculos nos estudos pedagógicos e avaliativos. A credibilidade da avaliação escolar residia na perfeição da construção de instrumentos de medida em busca de uma objetividade mensurável. Hoje, a identificação da essência do ato de avaliar com juízos de valor deixa claro que toda avaliação tem por referência um padrão que representa o valor vigente, dependente da questão cultural e, assim, variável no tempo e no espaço.

A avaliação como medida apresenta em seus aspectos ontológicos e epistemológicos alguns pressupostos distintos em relação à visão de mundo e de homem, e também de educação e avaliação. O homem é concebido nesse modelo como uma realidade pronta, objetiva e acabada, um ser permeado por verdades perenes e não sujeito a alterações. Bonniol e Vial (2001) esclarecem que esse tipo de avaliação já não satisfaz os avaliadores, porque apresenta uma concepção mecanicista do mundo e se inscreve na ideologia positivista, sobretudo porque tende a transformar em dogma a ideia da monocausalidade linear: a causalidade não é mais a explicação suficiente de um fenômeno. Compreender não é mais procurar a causa.

A constatação de que, nas situações de vida e nas práticas sociais, a explicação pode ser pluricausal e não-linear afetou a perenidade desse modelo e sugeriu a necessidade de outras Há uma busca por aparente simplificação do processo, voltando os olhares, prioritariamente, para os resultados observáveis, medidos e quantificados, pois apenas esses são considerados "científicos":

> Preferindo os produtos acabados por serem mais estáveis e favorecerem a medição e a comparação, reduzindo-se a um controle do resultado observado em relação à medida esperada – uma realidade multidimensional resumida a uma simples média mediante um vetor de coeficientes de ponderação. (BONNIOL; VIAL, 2001, p. 54)

De qualquer forma, podemos evitar uma apropriação acrítica e pouco reflexiva dos instrumentos e critérios adotados nos testes objetivos na avaliação de nossos sistemas educacionais:

> A relativização da importância da ideia de medida, da necessidade de uma organização linear do conhecimento, da existência de padrões universais a serem perseguidos, não minimiza a importância dos testes objetivos em processos de avaliação, mas obriga a um repensar sobre sua forma de utilização. Assim como medir ou não medir não parece ser mais a questão, usar ou não usar também já deixou de sê-lo. (MACHADO, 1996, p. 299)

Pontuo com relação a essa intenção inicial que não considero "um exame" como um instrumento, por mais perfeito que seja, que garantirá a qualidade da educação no país sem que repensemos nos inúmeros problemas sociais, estruturais e educacionais que o envolvem. Dessa forma, questiono se poderia um instrumento de avaliação assumir a responsabilidade por algo que não é capaz de prever ou transformar.

Sabe-se que o exame é um meio a serviço dos fins maiores do sistema educacional, e estes precisam ser claros, coerentes e revestir-se de força política e engajamento coletivo na sua realização. Em hipótese nenhuma deveriam subsidiar a construção de manchetes exageradas e sensacionalistas ou apoio para campanhas políticas e comerciais.

Retomo algumas questões que não podem ser silenciadas neste momento em que precisamos estar conscientes de que inúmeros esforços para medir com precisão os produtos educacionais não influenciarão diretamente a qualidade democrática da escola brasileira, e de que aspectos políticos e epistemológicos podem ser escondidos ou silenciados nesse processo.

No âmbito das discussões sobre a concepção do ato de medir e avaliar, talvez a ideia mais clara para responder às

arguições, neste momento, venha do professor Nílson Machado (1997) quando diz que "um processo de avaliação nunca se esgota em um processo de medida, porém vai além dele". Dessa forma, antes de excluir a ideia de medida no ideário educacional e dos estudos em avaliação, ou mesmo minimizar a importância dos testes objetivos em avaliações de larga escala, posso afirmar que a concepção de avaliação transcende a noção de medida, o que nos impõe um repensar sobre a atual forma de utilização das mensurações educacionais.

Se hoje chegamos à versão contemporânea da psicometria, com a Teoria da Resposta ao Item (TRI), é porque, desde os primeiros estudos da avaliação educacional, já havia o desejo de alcançar resultados precisos da medida do desenvolvimento de um estudante. E a TRI parece ser o que há de mais inovador nessa direção.

A credibilidade na cientificidade e precisão de um produto educacional objetivo e versátil faz com que estudiosos no mundo inteiro busquem, há quase duzentos anos, construir um perfeito instrumento de medida para chegar à máxima objetividade com resultados precisos e mensuráveis. Será isso possível se pensarmos em seres humanos complexos e multifacetados como nós? Essa é uma questão que levantarei mais à frente nas considerações sobre os pressupostos epistemológicos que garantem a confiança no modelo da Teoria da Resposta ao Item.

Ao procurar elencar e refletir sobre a TRI em estudos disponíveis na literatura educacional e científica, fiz o recorte de alguns pesquisadores estudados e verifiquei que todos têm se dedicado a essas questões de forma especial em nosso país. São eles: Dalton Andrade (2010; ANDRADE; VALLE, 1998; ANDRADE et al., 2000); Raquel da Cunha Valle (2000, 2001, 2002; ANDRADE; VALLE, 1998); Héliton Tavares (ANDRADE; TAVARES; VALLE, 2000); Ruben Klein (2009); Tufi Machado Soares (2005); Luiz Pasquali (2011) e Ronald Tarjino Nojosa (2002; 2010).

Na literatura internacional destaco ainda alguns estudos consultados, como os de Michael Kolen e Robert Brennan (2010); Dany Laveault e Jacques Grégoire (2010) e Christine Demars (2010).

PSICOMETRIA: DA TEORIA CLÁSSICA DOS TESTES À TEORIA DA RESPOSTA AO ITEM

No entendimento de Andrade e Valle (1998), resultados obtidos em provas, expressos apenas por seus escores brutos ou padronizados, têm sido frequentemente utilizados nos processos de avaliação e seleção de indivíduos.

No entanto, os resultados encontrados dependem do particular conjunto de itens que compõem o instrumento de medida, ou seja, as análises e interpretações estão sempre associadas à prova em sua totalidade, e essa é a característica principal da Teoria Clássica das Medidas. Assim, a comparação entre indivíduos somente é possível quando eles são submetidos às mesmas provas ou, pelo menos, ao que se denomina de provas paralelas:

> Atualmente, na área educacional, vem crescendo o interesse pela aplicação de técnicas derivadas da Teoria da Resposta ao Item (TRI) que propõem modelos de variáveis latentes para representar a relação entre a probabilidade de um aluno responder corretamente a um item e seus traços latentes ou habilidades na área do conhecimento avaliada, os quais não são observados diretamente. Tendo como elemento central os itens e não a prova como um todo, a TRI permite, por exemplo, a comparação entre populações distintas submetidas a provas diferentes, mas com alguns itens comuns ou, ainda, a comparação entre indivíduos da mesma população que tenham sido submetidos a diferentes provas, com ou sem itens comuns. (ANDRADE; VALLE, 1998, p. 13)

Os conceitos básicos da teoria psicométrica fundamentada no item tiveram início com os trabalhos de Lawley (1943) e foram posteriormente enriquecidos com os estudos desenvolvidos por Lord (1952). Assim, entre os anos 1950 e 1960, a TRI já buscava responder a indagações relativas aos testes de inteligência, cujos resultados variavam em razão dos instrumentos de medida utilizados. Qual seria, então, o resultado correto? E quando o objeto a ser medido era a inteligência humana? A solução dada pela TRI a esse problema – independência do instrumento de medida em relação ao objeto que se deseja medir – utilizava modelos e algoritmos matemáticos difíceis de serem operacionalizados à época.

Por isso, somente após o avanço tecnológico dos anos 1980, com o desenvolvimento de *softwares* para uso prático dos algoritmos complexos que o modelo contém, é que a metodologia começou a se expandir com crescente intensidade. Atualmente, a maioria dos programas de avaliação educacional em larga escala no mundo tem como base a Teoria da Resposta ao Item (BRASIL, 2012).

Pode-se destacar que a TRI é uma modelagem estatística de aplicação frequente em testes de conhecimento, e seu uso é consagrado na área de educação em vários países. O interesse por essa modelagem estatística vem crescendo entre os educadores e gestores brasileiros em virtude da sua aplicação em avaliações internacionais e nacionais em larga escala, inicialmente com o Sistema de Avaliação do Ensino Básico (Saeb), e nas avaliações regionais, como o Sistema de Avaliação do Rendimento Escolar do Estado de São Paulo (Saresp), nas quais a técnica vem sendo utilizada desde 1995 e 1996, respectivamente.

Existem diferentes técnicas para se obter uma medida do constructo, que aqui se refere, especialmente, ao desempenho acadêmico; grosso modo, podem ser divididas segundo as chamadas técnicas clássicas, cujo modelo para a construção da escala baseia-se diretamente no resultado obtido no instrumento globalmente.

Há, também, técnicas cujos modelos específicos são construídos para cada item do instrumento, sendo que a elaboração da escala considera todos esses modelos individuais.

Nesse último caso, podem ser classificadas as técnicas baseadas na Teoria de Resposta ao Item (TRI) que, originalmente, foram empregadas na produção de escalas de proficiência em testes de avaliação educacional (SOARES, 2005).

A Teoria da Resposta ao Item surgiu da necessidade de superar as limitações da apresentação de resultados somente por meio de percentuais de acertos ou escores dos testes e, ainda, da dificuldade de comparar resultados de diferentes testes, em diversas situações:

> Na Teoria Clássica dos Testes, os resultados dependem do particular conjunto de questões que compõem a prova e dos indivíduos que a fizeram, ou seja, as análises e interpretações estão sempre associadas à prova como um todo e ao grupo de indivíduos. Assim, a comparação entre indiví-

> duos ou grupos de indivíduos somente é possível quando eles são submetidos às mesmas provas ou, pelo menos, ao que se denomina de provas paralelas, quase sempre difíceis de serem construídas. Desta maneira, fica muito difícil fazer comparações quando diferentes indivíduos fazem provas diferentes. Isto é importante em avaliações de larga escala quando se quer avaliar uma grande parte de um currículo de uma determinada disciplina e série e para tal, é necessário apresentar um grande número de itens aos alunos, maior do que eles poderiam responder em uma ou duas horas de prova. Além disso, nas avaliações ao longo dos anos, torna-se difícil comparar resultados quando as provas não são as mesmas. A TRI muda o foco de análise da prova como um todo para a análise de cada item. (KLEIN, 2009, p. 126-127)

Sendo a TRI um conjunto de modelos matemáticos em que a probabilidade de resposta a um item é modelada como função da proficiência (habilidade) do aluno (variável latente, não observável) e de parâmetros que expressam certas propriedades dos itens, quanto maior a proficiência do aluno, maior a probabilidade de acertar o item.

A Teoria da Resposta ao Item tem sido conceituada como uma modelagem estatística de aplicação frequente em testes de conhecimento:

> Modelos construídos pela Teoria da Resposta ao Item (TRI) mostram a relação entre a capacidade ou traço latente (simbolizado pela letra grega *theta*) medido pelo instrumento e a resposta ao item.
>
> A resposta ao item pode ser dicotômica (duas categorias), tais como certo ou errado, sim ou não, concordar ou discordar. Ou, pode ser politômica (mais de duas categorias), tais como uma classificação de um juiz ou apontador [...]. O construto medido pelos itens podem ser de proficiência acadêmica ou aptidão, ou pode ser uma atitude ou crença. (DEMARS, 2010, p. 3)

Na TRI, não se pergunta quantos itens o sujeito acertou, e sim por que acertou ou errou cada item individual. Dessa forma, a TRI está interessada em descobrir qual é o tamanho

do *theta* que o sujeito deve ter para poder acertar cada item individualmente. Pode-se inferir, então, que, em teoria, basta até um único item para se poder descobrir o tamanho do *theta* do sujeito.

O valor do *theta*, normalmente, varia entre -3 e +3. Ele expressa a quantidade de habilidade que o sujeito deve ter para acertar o item. A TRI possui três modelos, que se distinguem pelo número de parâmetros utilizados para descrever o item. O modelo de um parâmetro avalia a dificuldade do item; o de dois parâmetros avalia a dificuldade e a discriminação; e o de três parâmetros avalia a dificuldade, a discriminação e a resposta correta dada ao acaso. Além dos parâmetros de discriminação e de dificuldade, algumas avaliações fazem uso de um parâmetro para controlar o acerto casual, como é o caso de nossa análise com o novo Exame Nacional do Ensino Médio (Enem). Este último parâmetro tem um papel bastante importante nas avaliações com itens de múltipla escolha.

Dentre as vantagens atribuídas à Teoria da Resposta ao Item, destacamos o fato de que esse instrumento da psicometria contemporânea permite superar certas limitações da psicometria tradicional, sobretudo questões que dependem diretamente da amostra de sujeitos utilizada na avaliação em larga escala. Se, nesse tipo de análise, a amostra não for rigorosamente representativa da população, alguns resultados não poderão ser considerados válidos.

PRESSUPOSTOS EPISTEMOLÓGICOS DA TRI

UNIDIMENSIONALIDADE

Kolen e Brenan, ao se referirem à suposição unidimensional do modelo da TRI, explicitam da seguinte forma o conceito:

> Modelos de Teoria da Resposta ao Item Unidimensionais de itens dicotômicos (0,1) em testes padronizados assumem que a capacidade do examinando é descrita por uma variável latente única, conhecida como teta e definida de modo que – oo <0< oo. A utilização de uma variável latente única implica que o construto a ser medido pelo teste é unidimensional. Em termos práticos a suposição de unidimensionalidade na TRI exige que os testes meçam apenas uma habilidade. Por

> exemplo, um teste de matemática que contém alguns itens que são estritamente computacionais e outro que envolve material verbal provavelmente não são unidimensionais. (2010, p. 156-157)

Um dos pressupostos centrais para a concretização e aplicabilidade real do modelo logístico unidimensional de três parâmetros é a

> [...] homogeneidade do conjunto de itens que supostamente devem estar medindo um único traço latente. Em outras palavras, deve haver apenas uma habilidade responsável pela realização de todos os itens da prova. (ANDRADE; TAVARES; VALLE, 2000, p. 16)

Talvez pela complexidade da afirmação que sustenta a existência do modelo, os próprios estatísticos e pesquisadores que declararam acima que um único traço ou habilidade pode ser medido registram outra sentença que parece abalar o pressuposto da unidimensionalidade imposto pela literatura especializada: "parece claro que qualquer desempenho humano é sempre multideterminado ou multimotivado, dado que mais de um traço latente entra na execução de qualquer tarefa" (ANDRADE; TAVARES; VALLE, 2000, p. 16).

Se mais de uma habilidade ou traço latente compõe cada desempenho humano, e se pesquisadores e cientistas estão cientes da complexidade que traduz o registro de aprendizagens tão variadas que o indivíduo constrói nos diferentes espaços de educação formal e não-formal, como uma avaliação pode se propor a medir um fator unidimensional em indivíduos com formação tão plural e multifacetada?

Para satisfazer o postulado é suficiente admitir que há uma habilidade dominante (um fator dominante) responsável pelo conjunto de itens ou questões do exame? Este fator é o que se supõe estar medindo com o teste?

> O modelo proposto pressupõe que o número de traços latentes medidos pela prova é igual a 1, isto é, o modelo supõe que a prova mede uma única habilidade. Tradicionalmente tem-se utilizado a técnica de análise fatorial com base na matriz de correlações tetracóricas para a verificação da dimensionalidade de provas. (ANDRADE; VALLE, 1998, p. 19)

E a literatura especializada segue reforçando o mesmo postulado que mantém o modelo em ação:

> Um teste que é unidimensional consiste de itens que abrangem uma única dimensão. Sempre que apenas uma pontuação é descrita em um teste, não está implícito que os itens partilham de construto primário comum [...]. Unidimensionalidade significa que um modelo possui um teta único para cada examinando e quaisquer outros fatores que afetam a resposta ao item são aleatórios ou erros de um único item e não afetam os demais. (DEMARS, 2010, p. 32)

No entanto, o estatuto de que a habilidade e a cognição humanas e suas diferentes manifestações não podem ser facilmente medidas remonta aos primórdios da psicometria nos séculos XVIII e início do século XIX:

> [...] a ideia de que um conceito tão impreciso e tão dependente do contexto social como a inteligência pode ser identificado como uma "coisa" única localizada no cérebro e dotada de um determinado grau de hereditariedade, e que, portanto, pode ser medida e receber um valor numérico específico que permite uma classificação unilinear das pessoas em função da quantidade de inteligência que cada um supostamente possui. Ao identificar um eixo fatorial matemático com o conceito de "inteligência geral", Spearman e Burt forneceram uma justificação teórica da escala unilinear que Binet havia proposto como simples guia aproximativo. (GOULD, 2003, p. 252)

O psicólogo Luiz Pasquali (2011), ao se referir aos pressupostos da TRI, também afirma que há um problema a ser enfrentado na questão da representação comportamental. Afinal, qual seria a maneira adequada de representar e observar empiricamente atributos latentes para que possam ser cientificamente abordados?

> O comportamento humano tipicamente se apresenta como multimotivado, dado que fatores múltiplos entram na sua aparição, sendo, portanto, difícil, se não impossível, determinar causas ou fatores únicos para qualquer comportamento, ao menos de adultos. Isto implica que seria impossível, determinar causas ou fatores únicos para qualquer comportamento, ao menos de adultos.

> Isto implica que seria impossível definir comportamentos (itens) críticos para qualquer traço latente, no sentido de um comportamento "x" ser específico e único de tal traço e não interfaceando com qualquer outro traço. (PASQUALI, 2011, p. 63-64)

No livro *Introduction aux théories des tests en psychologie et en sciences de l'éducation*, os pesquisadores Laveault e Grégoire explicitam que a exigência da unidimensionalidade

> [...] significa que todos os itens de um teste devem medir um único traço. Na prática, esse critério nunca é plenamente cumprido devido a inevitáveis erros de medição e da complexidade dos traços medidos. (2010, p. 279)

Se pensarmos em um exame específico utilizado em larga escala em nosso país, como o Enem, podemos assegurar que a TRI baseada em um modelo logístico unidimensional de três parâmetros é a melhor opção metodológica para esse caso?

Se os modelos unidimensionais possuem matrizes gigantescas para a elaboração e tratamento de dados e requerem programas sofisticadíssimos, como poderíamos pensar em modelos multidimensionais, que analisam mais de um traço latente, quando estes, embora já façam parte dos estudos teóricos, ainda são de difícil execução na prática?

Com essas inquietações em mente, em 2002, Nojosa publicou na revista *Estudos em Avaliação Educacional*, da Fundação Carlos Chagas, um estudo inicial que considerou pioneiro por buscar conhecer melhor o funcionamento da TRI multidimensional. A intenção era avaliar sob esse novo enfoque os resultados do Exame Nacional do Ensino Médio de 1999, que, por ser um exame interdisciplinar, sugeria a necessidade de utilização de um modelo em que a habilidade fosse expressa por mais de uma dimensão. O autor já pontuava dez anos antes que a importância do Enem no contexto nacional estava motivando o estudo dos modelos multidimensionais da TRI.

O grande diferencial desse exame pode ser atribuído aos itens interdisciplinares que o compõem, pois cada um deles foi construído de modo a avaliar até cinco competências, considerando a primeira versão do Enem. O autor afirma que modelos unidimensionais de TRI não se aplicam ao Enem, que foi estruturado segundo uma matriz de cinco competências. Ele sustenta

que alguns exames, seja pela construção dos itens, seja pela própria finalidade da avaliação, não podem ser considerados unidimensionais, como, por exemplo, o próprio Enem:

> Estatísticas que descrevem características de itens são comumente empregadas no processo de construção de testes. Estas estatísticas são frequentemente usadas para produzir formas equivalentes de testes ou para produzir testes com características específicas. De modo geral, essas estatísticas assumem que o item está medindo uma única habilidade. Entretanto, os itens são, geralmente, multidimensionais em algum sentido e, dependendo da intensidade das dimensões, as estatísticas unidimensionais não são apropriadas. Alguns itens medem ou exigem de forma mais dominante uma só habilidade. Para estes itens as estatísticas unidimensionais são razoáveis. Por outro lado, itens que requerem claramente mais de uma habilidade necessitam de um tratamento diferenciado, ou seja, necessitam de medidas que levem em consideração as diferentes dimensões da habilidade. (NOJOSA, 2002, p. 145)

Na expressão de Nojosa (2002), a violação do pressuposto da unidimensionalidade conduz a consequências negativas como a desconfiança na própria validade do item construído e a impossibilidade de se garantir a independência local quando se assume um modelo unidimensional com itens multidimensionais.

INDEPENDÊNCIA LOCAL

Outra suposição do modelo, a chamada independência local ou independência condicional, é que, para uma dada habilidade, as respostas aos diferentes itens da prova são independentes. Essa suposição é essencial para o processo de estimação dos parâmetros do modelo. Na realidade, como a unidimensionalidade implica independência local, tem-se somente uma e não duas suposições a serem verificadas.

Assim, itens devem ser elaborados de modo a satisfazer a suposição de unidimensionalidade:

> [...] as respostas aos itens não são localmente independentes, sob um modelo unidimensional uma outra dimensão deve estar causando a dependência. Com independência local nos testes o foco está na dependência entre os pares de itens. (DEMARS, 2010, p. 48)

Kolen e Brenan definem essa suposição da seguinte maneira:

> Independência local significa que depois de levar em conta a capacidade do examinando, observamos que as respostas aos itens são estatisticamente independentes. Sob o efeito da independência local a probabilidade de que os examinandos com determinada habilidade respondam corretamente tanto o item 1 quanto o item 2 é igual a probabilidade de responder corretamente ao item 1 e responder corretamente ao item 2. A suposição de independência local implica que não existam dependências entre outros itens que aqueles que são atribuídos á capacidade latente. Um exemplo em que a independência local provavelmente não se concretizaria é quando os testes são compostos por conjuntos de itens que são baseados em estímulos comuns como a leitura de passagens e gráficos. Nesse caso a independência local provavelmente seria violada [...]. (2010, p. 157)

A exigência de independência local significa que o traço que é o objeto da avaliação deve ser o único fator que determina a variabilidade das respostas aos itens de um teste. Uma vez que a característica medida foi considerada, nenhuma relação deve existir entre as respostas de um assunto em diferentes itens.

Se as instruções de um teste fornecerem pistas para alguns itens diferentes, a exigência de independência local não mais terá sido respeitada. Nesse caso, o resultado do teste vai depender não só do traço a ser medido, mas também da capacidade para encontrar algumas pistas úteis.

Se um item de matemática, por exemplo, pede algum conhecimento especial em geografia, o sucesso desse item não dependerá exclusivamente do traço latente que queremos medir (LAVEAULT; GRÉGOIRE, 2010). A questão da excessiva fragmentação das 120 habilidades na matriz que compõe o Novo Enem[2] torna-se um obstáculo ao pressuposto da independência local, parecendo uma tarefa quase impossível elaborar questões totalmente independentes em um único exame de 180 itens.

QUALIDADE E ADEQUAÇÃO DOS ITENS

Um pressuposto importante de qualquer modelo para avaliação educacional que garanta a comparabilidade dos resultados é que o item apresente o mesmo comportamento nos

2 O Novo Enem foi reformulado em 2009 como uma prova interdisciplinar estruturada em torno de 4 matrizes: Linguagens, Código e suas Tecnologias; Matemática e suas Tecnologias; Ciências da Natureza e suas Tecnologias e Ciências Humanas e suas Tecnologias. Cada matriz do Novo Enem prevê que o aluno, a partir dos eixos cognitivos estabelecidos, trabalhe com 30 competências distribuídas pelas 4 áreas do conhecimento acima descritas, sendo que essas competências se subdividem em 120 habilidades que nortearão a elaboração das 180 questões de múltipla escolha do exame.

diversos grupos populacionais que estão sendo avaliados. Quando se utilizam os modelos da TRI, essa questão do comportamento se traduz na estabilidade dos parâmetros dos modelos dos itens para as diferentes populações. No entanto, embora em grau elevado o comportamento diferencial do item (DIF) possa prejudicar a comparabilidade dos resultados, quando moderado e localizado em poucos itens, o DIF além de praticamente não afetar a proficiência produzida, pode, se devidamente analisado, trazer informações importantes sobre diferenças curriculares e diferenças socioculturais, por exemplo, entre as regiões.

Um banco de itens com uma grande quantidade de questões bem elaboradas e adequadas é condição essencial para o correto desenvolvimento e aplicação da Teoria da Resposta ao Item. Moreira Júnior define o Banco de Itens (BI) como:

> [...] uma base de dados de itens formada por uma parte descritiva (enunciado, opção correta, opções incorretas), uma parte de informação psicométrica (parâmetros estimados dos itens, tanto os da TCT quanto os da TRI) e qualquer outra informação relevante (por exemplo, conteúdo que cada item mede, dificuldade teórica, taxas de exposição do item). (2011, p. 61)

Todas as questões utilizadas em uma edição dos exames nacionais do ensino médio precisam ser pré-testadas e analisadas para que possam compor adequadamente o banco de itens, sendo classificadas como fáceis, médias e difíceis. A logística dessa pré-testagem não é simples e os dados devem ser mantidos em absoluto sigilo:

> Uma avaliação, qualquer que seja a sua natureza, demanda a pré-testagem dos instrumentos, a fim de adequá-los aos sujeitos integrantes do conjunto avaliado. Isso, naturalmente, exige que se tenha uma amostra representativa, o que em muitas situações constitui-se num quadro bastante problemático por não atender a princípios definidores, dando margem a discussões, com o envolvimento do grande público e o comprometimento da validade de todo o processo, que passa a não merecer a credibilidade da sociedade. (VIANNA, 2005, p. 133)

Depois dos pré-testes, as questões podem ser eliminadas, reformuladas ou incorporadas a um banco de itens, que deve ser constantemente atualizado. Testadas antes da prova, as questões ganham um peso que varia de acordo com o desempenho dos estudantes nos pré-testes, e quanto mais alunos acertarem uma determinada pergunta, menor o peso que ela terá na prova, porque o grau de dificuldade é supostamente menor. A proposta do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) é construir um banco com milhares de itens para ter a capacidade de fazer inúmeras e distintas provas com as mesmas possibilidades de comparação.

A Teoria da Resposta ao Item como metodologia para a construção e análise de avaliações em larga escala permite a construção de escalas gigantescas que agregam matrizes capazes de conter, na mesma métrica, milhares de itens e proficiências de diferentes indivíduos em distintos momentos e testes, permitindo a associação e comparabilidade dos dados ali dispostos. Agora, a confiabilidade dessa escala depende diretamente da qualidade dos itens ou questões de cada prova realizada (ou seja, envolvendo altos níveis de discriminação e diferentes níveis de dificuldade). Isso significa que problemas na construção de itens comprometem decisivamente o poder de equalização da prova, que é seu mérito de poder equiparar, tornar comparáveis diferentes itens e diferentes indivíduos ao longo do tempo, sendo essa uma das principais vantagens apontadas para o uso da TRI na construção e análise de avaliações em larga escala. É essencial que se tenha um adequado e confiável banco de itens.

Nas discussões atuais presentes na mídia brasileira e nos editais do Inep em 2011 e 2012, acompanhamos a preocupação do Ministério da Educação com a ampliação do banco de itens em curto espaço de tempo para garantir a possibilidade de realização do exame em diferentes edições anuais.

Reforço, enfaticamente, que, até construirmos um banco de itens adequado de forma quantitativa e qualitativa, é essencial a composição de uma banca transitória de especialistas que se responsabilizem pela organização de toda a prova. Até mesmo a Constituição de um país requer disposições transitórias. Quanto mais um exame destinado a mais de 5 milhões de estudantes.

Dessa forma, há maior tempo para se buscar uma consciência da complexidade desse processo, evitando caminhos aligeirados que induzam à fraude e à corrupção. O período transitório entre a "banca e o banco" é a melhor possibilidade que essa investigação pode encaminhar às equipes responsáveis e agentes educacionais.

Considero a questão da unidimensionalidade um dos grandes desafios apresentados para a coerência epistemológica e prática do modelo proposto. Como lidar com a ideia unidimensional em face de um ser humano altamente complexo, objeto dos instrumentos de avaliação, e que no desempenho de qualquer tarefa mobiliza mais de um traço latente? Como afirmar que apenas um traço, construto ou habilidade estará sendo medido por um conjunto de itens? E não são os filósofos e educadores que levantam, em primeira instância, esses questionamentos e incongruências, e sim os próprios estatísticos e pesquisadores da área, como Andrade, Tavares e Valle (2000), Demars (2010), Nojosa (2002), Gould (2003), Laveault e Grégoire (2010) e Pasquali (2011).

Por que, então, políticos e administradores educacionais se referem à utilização da Teoria de Resposta ao Item em nossas avaliações como garantia de total confiabilidade para a testagem empregando um discurso superficial e desinformado em entrevistas para a mídia jornalística e televisiva, sem uma consideração, reflexão e estudo mais aprofundado? Poderemos, como leigos, declarar o que os próprios especialistas não afirmam?

Os modelos multidimensionais da Teoria da Resposta ao Item ou a Teoria da Resposta ao Item Multidimensional (Trim) poderiam ser uma resposta para esse desafio. Mas isso ocorrerá apenas quando os modelos multidimensionais deixarem o estatuto de "teoria" e forem aplicáveis na prática, o que, no momento, parece irrealizável para a engenharia computacional. Não duvidemos, no entanto, do possível desenvolvimento científico na área.

REFERÊNCIAS

ANDRADE, Dalton Francisco de. A Teoria da Resposta ao Item (TRI). *Avalia em ação*: ensinar com qualidade e valores, São Paulo, n. 3, p. 26-27, 2010.

ANDRADE, Dalton Francisco de; TAVARES, Héliton Ribeiro; VALLE, Raquel da Cunha. *Teoria da Resposta ao Item*: conceitos e aplicações. São Paulo: Associação Brasileira de Estatística, 2000.

ANDRADE, Dalton Francisco de; VALLE, Raquel da Cunha. Introdução à teoria da resposta ao item: conceitos e aplicações. *Estudos em Avaliação Educacional*, São Paulo, n. 18, p. 13-32, 1998.

BARRIGA, Angel Díaz. Uma polêmica em relação ao exame. In: ESTEBAN, Maria Teresa (Org.). *Avaliação*: uma prática em busca de novos sentidos. 4. ed. Rio de Janeiro: DP&A, 2003.

BONNIOL, Jean-Jacques; VIAL, Michel. *Modelos de avaliação*. Porto Alegre: Artmed, 2001.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Primeira oficina do Banco Nacional de Itens de 2012* acontece de 27/02 a 2/3. Disponível em: <http://portal.inep.gov.br/visualizar//asset_publisher/6AhJ/content/primeira-oficina-de-2012-do-banco-nacional-de-itens-do-enem-acontece-de-27-2-a-2-3>. Acesso em: 26 mar. 2012.

DEMARS, Christine. *Item Response Theory*. New York: Oxford University, 2010. Understanding Statistics.

GATTI, Bernardete A. Avaliação de sistemas educacionais no Brasil. *Sísifo*: Revista de Ciências da Educação, Lisboa, n. 9, p. 7-18, 2009. Disponível em: <http://sisifo.fpce.ul.pt>. Acesso em: jan. 2010.

GOULD, Stephen Jay. *A Falsa medida do homem*. São Paulo: Martins Fontes, 2003.

HAMBLETON, Ronald K.; SWAMINATHAN, Hariharan; ROGERS, H. Jane. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage, 1991. (Measurement Methods for the Social Sciences, v. 2).

KLEIN, Ruben. Utilização da Teoria da Resposta ao Item no Sistema Nacional de Avaliação da Educação Básica – Saeb. *Meta-Avaliação*, Rio de Janeiro, v. 1, n. 2, p. 125-140, maio/ago. 2009.

KOLEN, Michael J.; BRENNAN, Robert L. *Test Equating, Scaling, and Linking*: methods and practices. Iowa City, USA: Springer, 2010.

LAVEAULT, Dany; GREGÓIRE, Jacques. *Introduction aux théories des tests en psicologie et en sciences de l'éducation*. 2. ed. Bruxelles: De Boeck, 2010. (Méthodes en Sciences Humaines).

LAWLEY, D. N. On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh*, v. 61A, p. 273-287, 1943.

LORD, Frederic M. Theory of test scores. *Psychometric Monograph*, n. 7, 1952.

MACHADO, Nilson. *Epistemologia e didática*. São Paulo: Cortez, 1996.

_____. *Cidadania e educação*. São Paulo: Escrituras, 1997.

_____. *Educação, projetos e valores*. São Paulo: Escrituras, 2006.

MOREIRA JUNIOR, Fernando de Jesus. *Sistemática para implantação de testes adaptativos informatizados baseados na Teoria da Resposta ao Item*. 334 f. Tese

(Doutorado em Engenharia de Produção) – Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2011.

NOJOSA, Ronald Targino. Teoria da Resposta ao Item (TRI) – modelos multidimensionais. *Estudos em Avaliação Educacional*, São Paulo, n. 25, p. 123-166, jan./jun. 2002.

_____. *Interferência Bayesiana em modelos multidimensionais de resposta do item*. Tese (Doutorado) – Instituto de Matemática e Estatística (IME), Universidade de São Paulo, 2010.

PASQUALI, Luiz. *Psicometri*a: teoria dos testes na psicologia e na educação. 4. ed. Petrópolis, RJ: Vozes, 2011.

SOARES, Tufi Machado. Utilização da teoria da resposta ao item na produção de indicadores sócio-econômicos. *Pesqui. Oper*., v. 25, n. 1, p. 83-112, 2005.

VALLE, Raquel da Cunha. Teoria da Resposta ao Item. *Estudos em Avaliação Educacional*, São Paulo, n. 21, p. 7-91, jan./jun. 2000.

_____. A Construção e a interpretação das escalas de conhecimento – considerações gerais e uma visão do que vem sendo feito no Saresp. *Estudos em Avaliação Educacional*, n. 23, p. 71-92, jan./jun. 2001.

_____. Comportamento Diferencial do Item – (DIF): uma apresentação. *Estudos em Avaliação Educacional*, São Paulo, n. 25, p. 167-183, jan./jun. 2002.

VIANNA, Heraldo Marelim. *Fundamentos de um programa de avaliação educacional*. Brasília: Liber Livro, 2005.

CRISTINA ZUKOWSKY TAVARES

Pós-doutora em Educação pela Faculdade de Educação da Universidade de São Paulo (FE/USP). Coordenadora de Extensão Universitária. Docente do Curso de Mestrado em Promoção da Saúde do Centro Universitário Adventista de São Paulo (UNASP)
cristina.neri@unasp.edu.br

Recebido em: SETEMBRO 2012
Aprovado para publicação em: JANEIRO 2013