APLICAÇÃO DA TEORIA DA RESPOSTA AO ITEM UNI E MULTIDIMENSIONAL

PEDRO A. BARBETTA
LIGIA M. V. TREVISAN
HELITON TAVARES
TÂNIA C. ARANTES DE MACEDO AZEVEDO

RESUMO

Este trabalho mostra a aplicação de modelos da teoria da resposta ao item unidimensional e multidimensional numa prova multi e interdisciplinar usada em um processo seletivo de ingresso na universidade. Mesmo sendo esta uma prova composta por itens das diversas disciplinas do ensino médio, foi possível ajustar um modelo unidimensional da teoria da resposta ao item e construir uma escala pedagogicamente interpretável. Entretanto, com a abordagem multidimensional, foi possível identificar três traços latentes predominantes: raciocínio lógico, compreensão de texto e proficiência em inglês. O artigo relata também alguns ensaios realizados com o posicionamento e avaliação de itens no plano formado pelos traços latentes raciocínio lógico e compreensão de texto, assuntos ainda incipientes nos estudos de avaliação educacional.

PALAVRAS-CHAVE TEORIA DA RESPOSTA AO ITEM • DIMENSIONALIDADE • AVALIAÇÃO EDUCACIONAL • VESTIBULAR.

RESUMEN

Este trabajo muestra la aplicación de modelos de la teoría de la respuesta al ítem unidimensional y multidimensional en un test multi e interdisciplinario utilizado en un proceso selectivo de ingreso a la universidad. Aunque el test estuviera compuesto de ítems de las diversas asignaturas de la educación media, fue posible ajustar un modelo unidimensional de la teoría de la respuesta al ítem y construir una escala pedagógicamente interpretable. Sin embargo, con el abordaje multidimensional, fue posible identificar tres rasgos latentes predominantes: razonamiento lógico, comprensión de texto y dominio en idioma inglés. El artículo relata asimismo algunos ensayos realizados con la posición y evaluación de ítems en el plano formado por los trazos latentes razonamiento lógico y comprensión de texto, temas todavía incipientes en los estudios de evaluación educacional.

PALABRAS CLAVE TEORÍA DE LA RESPUESTA AL ÍTEM • DIMENSIONALIDAD • EVALUACIÓN EDUCACIONAL • EXAMEN DE INGRESO A LA UNIVERSIDAD.

ABSTRACT

This paper presents the application of unidimensional and multidimensional item response theory models on a multi- and interdisciplinary test that was part of a college admission process. Even though the test comprised contents from the many high school subjects, we have managed to set up a unidimensional model of the item response theory and establish a pedagogically sound scale. However, by using the multidimensional approach, we have managed to identify three predominant latent traits: logical reasoning, reading comprehension and proficiency in English. This paper also reports on some trials carried out with the positioning and the assessment of items in the intersection of the logical reasoning and reading comprehension latent traits, both still recent additions to the educational assessment studies.

KEYWORDS ITEM RESPONSE THEORY • DIMENSIONALITY • EDUCATIONAL ASSESSMENT • ADMISSION TEST.

INTRODUÇÃO

As provas são normalmente construídas com a finalidade de avaliar algum traço latente dos avaliados. Uma prova de matemática, por exemplo, costuma ser construída para avaliar a proficiência do avaliado em matemática; uma prova de português é feita para avaliar a proficiência em português; e assim por diante. Esse processo é usado em grande parte das avaliações com propósito de seleção, sendo que o escore do avaliado é calculado em cada área ou disciplina por meio da soma dos pontos (teoria clássica) ou pela teoria da resposta ao item (TRI). Essa última permite interpretar pedagogicamente a escala, além de possibilitar comparações entre diferentes edições das provas, desde que haja itens comuns entre elas ou um mesmo conjunto de respondentes que tenham participado dessas diferentes edições.

Outros instrumentos de avaliação buscam avaliar competências e habilidades mais gerais dos examinados, contendo itens que exigem, além de conhecimento específico, a capacidade de compreensão de textos e o raciocínio lógico. É o caso do Pisa, do Enem realizado de 1998 a 2008 e das

provas de conhecimentos gerais que vêm sendo aplicadas desde 2009 aos candidatos do vestibular da Unesp. A questão que se coloca é: qual o significado de aplicação da TRI nesses casos, e o que se agrega ao se aplicarem modelos TRI multidimensionais (TRIM), em termos de informação sobre a pertinência da prova frente aos objetivos que orientaram a sua elaboração.

Na prática, os instrumentos baseados em itens, em geral, não são puramente unidimensionais, assim como os indivíduos devem ter múltiplas habilidades, as quais podem ser captadas se o instrumento contém quantidade razoável de itens correlacionados com essas habilidades. Modelos multidimensionais são úteis para definir as feições essenciais de instrumentos, em termos de dimensões que eles permitem aferir e, em consequência, delinear o perfil intelectual do examinado.

A utilização da TRI na avaliação educacional já está bastante consolidada. No Brasil, é realizada em importantes avaliações, como no Enem, na Prova Brasil, no Saresp etc., e muitos estudos técnicos têm sido realizados para verificar a adequação da aplicação da TRI em avaliações educacionais (PRIMI et al., 2013; BORGATTO et al., 2011; CHILDS; OPPLER, 2000). Estudos sobre a aplicação da abordagem multidimensional – TRIM –, por sua vez, usualmente restringem-se ao estudo da dimensionalidade, considerando sua relação com a análise fatorial clássica (WIRTH; EDWARDS, 2007; VITÓRIA et al., 2006; FUNDAÇÃO CESGRANRIO, 2004; LAROS et al., 2000).

O presente artigo explora a prova de conhecimentos gerais do vestibular da Unesp de 2011, que é uma prova multidisciplinar e interdisciplinar com 90 itens de múltipla escolha (FUNDAÇÃO VUNESP, 2011). Essa prova tem função classificatória para uma segunda etapa de avaliação. Neste trabalho, a prova é analisada, primeiramente, numa abordagem unidimensional, através da TRI; depois é feita uma análise da dimensionalidade; e em seguida é usada uma abordagem multidimensional.

METODOLOGIA

Normalmente, a TRI é aplicada em provas destinadas a medir um traço latente dos avaliados, que supostamente representa uma competência específica. O modelo de TRI mais adotado em avaliação educacional é o logístico de três parâmetros (ANDRADE et al., 2000; AYALA, 2009). Por esse modelo, a probabilidade de um avaliado *j*, com proficiência Θ_j, acertar o item *i* é dada por:

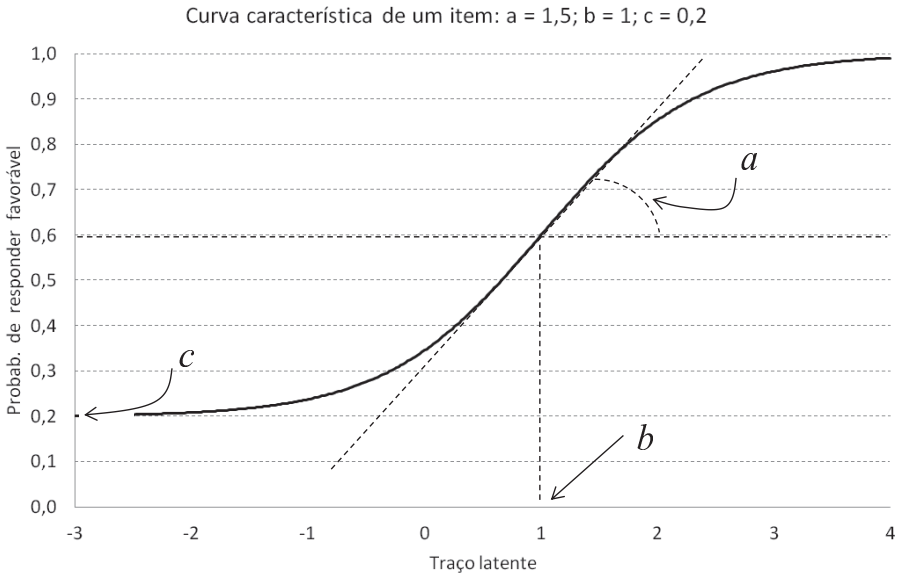
$$p_{ij} = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\Theta_j - b_i)}}$$

sendo que os parâmetros *a*, *b* e *c* são relativos ao item; e o parâmetro Θ está associado ao avaliado. Mais especificamente:

a_i representa a discriminação do item *i*;
b_i o nível de dificuldade do item *i*;
c_i a probabilidade de acerto casual do item *i*; e
Θ_j o traço latente do avaliado *j*.

A escala de medida do traço latente Θ tem, usualmente, média 0 e desvio padrão 1, seguindo uma distribuição normal. O parâmetro de dificuldade do item *b* é expresso na mesma escala de Θ. Assim, um item com $b = 2$ pode ser considerado difícil e um item com $b = -2$ fácil para um avaliado com proficiência mediana; da mesma forma, um indivíduo com $\Theta = 2$ tem alta proficiência e um indivíduo com $\Theta = -2$ baixa proficiência. O nível de discriminação do item *a* deve ser positivo, idealmente maior que 0,7. Quanto maior *a*, maior o nível de discriminação do item, porém, não é realista $a > 3$ (ver detalhes em ANDRADE et al., 2000; AYALA, 2009). A Figura 1 representa, geometricamente, esses parâmetros para um item hipotético.

FIGURA 1 – Curva representando a probabilidade de acerto de um item em função do traço latente θ



Fonte: Elaboração dos autores.

Os parâmetros a_i, b_i e c_i dos itens e os parâmetros θ_j dos avaliados são estimados objetivamente por meio de métodos estatísticos a partir das respostas dos avaliados e do modelo proposto. Com essas estimativas, os itens podem ser posicionados na escala θ, permitindo realizar uma interpretação pedagógica da mesma.

Há duas suposições básicas para a aplicação dos modelos usuais da TRI: unidimensionalidade e independência local. Por unidimensionalidade entende-se que a prova esteja medindo um traço latente único, que pode representar uma proficiência, ou mesmo uma composição de habilidades e proficiências dos avaliados. Por independência local entende-se que a dependência entre os itens é perfeitamente explicada pelo traço latente θ dos avaliados.

Numa prova multidimensional, em que há itens associados a várias habilidades ou proficiências, pode-se ter uma dimensão dominante que reflita alguma composição dessas habilidades ou proficiências presentes nos avaliados. Segundo Reckase (2009, p. 126), o parâmetro θ do modelo

unidimensional da TRI pode representar uma composição de habilidades ou proficiências. Esta capacidade da TRI de captar uma composição de proficiências também foi verificada por Barbetta *et al*. (2011), analisando as respostas de um teste composto por itens de matemática e itens de linguagens e códigos de provas do Enem. Outras formas de construir uma medida única em instrumentos multidimensionais também são discutidas em Reckase (2009) e Diaz (2010).

Num teste multidisciplinar, como o da primeira fase do vestibular da Unesp, é interessante ter uma medida única do avaliado, inclusive para atender ao requisito da classificação para a segunda fase do vestibular. No entanto, explorar a dimensionalidade da prova, identificando as habilidades ou proficiências que esta esteja medindo, produz novos conhecimentos sobre a prova e sobre os avaliados, permitindo um aprimoramento em futuras avaliações. Nesse contexto, a aplicação de modelos TRI multidimensionais (TRIM) torna-se bastante útil.

Há várias propostas de modelos TRIM, mas a mais comum considera que existem vários traços latentes, representando diferentes habilidades ou proficiências dos avaliados, mas apenas um parâmetro de dificuldade. Nesse modelo, a probabilidade de um avaliado *j*, com traços latentes Θ_{1j}, Θ_{2j}, ...Θ_{Kj}, acertar um item *i* é dada por:

$$p_{ij} = c_i + (1 - c_i)\frac{1}{1 + e^{-(a_{1i}\Theta_{1j} + a_{2i}\Theta_{2j} + \ldots a_{ki}\Theta_{kj} + d_i)}}$$

sendo que o parâmetro de dificuldade do item *i* pode ser definido por:

$$B_i = \frac{-d_i}{\sqrt{(a_{1i}^2 + a_{2i}^2 + \cdots a_{Ki}^2)}}$$

Os modelos de TRIM têm relação bastante forte com a análise fatorial, técnica clássica usada no estudo da dimensionalidade de instrumentos. As chamadas cargas fatoriais (correlações entre itens e fatores) podem ser obtidas com base nos parâmetros de discriminação dos itens. Com a relação

entre TRIM e análise fatorial, torna-se possível usar também as estratégias clássicas de análise fatorial na análise da dimensionalidade de provas e outros instrumentos. A análise fatorial feita com modelos de TRIM é conhecida como análise fatorial de informação completa (WIRTH; EDWARDS, 2007; BOCK; GIBBONS; MURAKI, 1988).

Não existe uma forma padrão para se verificar a dimensionalidade adequada do modelo, mas algumas técnicas podem orientar para um número adequado de traços latentes a serem incluídos no modelo, tais como a análise de componentes principais e a análise paralela sobre a matriz de correlação tetracórica dos itens, o teste qui-quadrado entre modelos com diferentes dimensões, entre outras (RECKASE, 2009).

Assim como na abordagem unidimensional, em que é possível posicionar itens e avaliados numa mesma escala, no caso bidimensional (provas avaliando as proficiências θ_1 e θ_2 dos avaliados) é possível posicionar itens e avaliados no plano formado pelos eixos cartesianos θ_1 e θ_2, explorando relações entre itens e avaliados em termos dos traços latentes θ_1 e θ_2.

As análises foram feitas com as respostas da prova de conhecimentos gerais do vestibular da Unesp (FUNDAÇÃO VUNESP, 2011). A calibração do modelo unidimensional foi realizada com toda a população (73.178 avaliados). Já a análise da dimensionalidade e os ajustes dos modelos multidimensionais foram feitas com uma amostra aleatória de 20.000 avaliados que fizeram pelo menos 20 pontos no total de 90. Na calibração dos modelos multidimensionais, foram usadas as estimativas dos parâmetros de acerto casual, c_i, da análise unidimensional.

Em termos computacionais, a análise unidimensional foi realizada com o *software* Bilog-MG (www.ssicentral.com) e os modelos multidimensionais com o pacote *mirt* (CHALMERS, 2012) do *software* livre R (R CORE TEAM, 2013). A função *mirt* desse pacote baseia-se nos princípios desenvolvidos por Bock e Aitkin (1981) e Bock, Gibbons e Muraki (1988).

ANÁLISE DA PROVA COM MODELO UNIDIMENSIONAL

A prova de conhecimentos gerais do vestibular da Unesp, apresentada em Fundação Vunesp (2011), é multidisciplinar. Mesmo assim, foi adotado, primeiramente, um modelo unidimensional de TRI com três parâmetros, que se ajustou bem às respostas dos avaliados. A Tabela 1 apresenta as estimativas dos parâmetros de discriminação e dificuldade dos itens, sendo θ fixado na escala de média zero e variância um. O parâmetro *c* não é apresentado por ter menos importância na presente discussão.

TABELA 1 – Estimativas dos parâmetros de discriminação (*a*) e de dificuldade (*b*) na prova de conhecimentos gerais do vestibular da Unesp, 2011

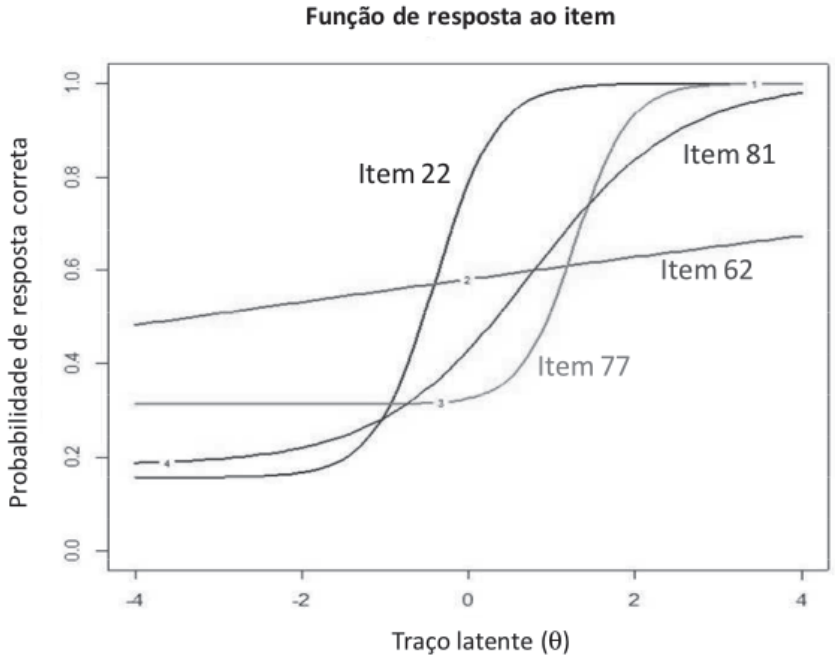
Item	*a*	*b*	Item	*a*	*b*	Item	*a*	*b*	Item	*a*	*b*	Item	*a*	*b*	Item	*a*	*b*
1	1,13	-1,81	**16**	0,68	-0,28	**31**	0,65	2,57	**46**	1,30	0,57	**61**	2,30	1,49	**76**	2,41	1,10
2	0,63	0,97	**17**	1,05	-0,18	**32**	0,25	0,17	**47**	1,31	0,22	**62**	0,17	0,76	**77**	3,17	1,27
3	1,37	1,73	**18**	1,22	1,84	**33**	1,00	-0,52	**48**	0,85	0,81	**63**	0,85	1,71	**78**	1,55	1,02
4	1,09	-0,04	**19**	1,03	-1,49	**34**	1,34	-1,05	**49**	1,90	-0,30	**64**	1,55	1,08	**79**	1,60	1,28
5	0,71	0,99	**20**	0,93	-0,23	**35**	1,45	1,67	**50**	3,10	2,83	**65**	1,58	1,89	**80**	1,42	0,94
6	1,49	-1,99	**21**	2,31	-0,86	**36**	1,79	0,36	**51**	1,33	-1,63	**66**	1,55	0,30	**81**	1,10	0,73
7	1,72	-1,59	**22**	2,72	-0,38	**37**	0,75	0,09	**52**	1,35	-0,49	**67**	1,35	-0,16	**82**	2,79	1,18
8	1,21	-1,47	**23**	1,28	0,26	**38**	1,15	-0,14	**53**	1,23	0,88	**68**	0,65	0,20	**83**	2,20	0,67
9	1,58	-0,32	**24**	2,63	0,00	**39**	1,41	-0,06	**54**	1,13	-0,75	**69**	0,73	0,60	**84**	3,54	2,24
10	1,25	-2,41	**25**	2,00	0,40	**40**	1,44	1,83	**55**	0,94	-0,66	**70**	1,32	0,82	**85**	1,68	1,46
11	1,56	0,60	**26**	1,19	1,40	**41**	0,96	-0,62	**56**	1,75	-0,56	**71**	2,48	1,14	**86**	1,97	0,91
12	1,38	0,22	**27**	1,32	-0,62	**42**	0,26	-0,40	**57**	0,92	-2,33	**72**	1,76	0,88	**87**	2,54	1,36
13	0,71	-2,12	**28**	2,05	0,32	**43**	0,68	-2,25	**58**	1,56	0,83	**73**	1,34	1,50	**88**	1,58	2,17
14	1,01	-0,75	**29**	1,15	0,42	**44**	1,03	-1,93	**59**	1,38	0,70	**74**	1,96	1,45	**89**	1,72	1,38
15	0,91	-1,55	**30**	1,46	1,22	**45**	1,72	0,75	**60**	1,47	-0,18	**75**	1,50	1,36	**90**	2,43	1,48

Fonte: Elaboração dos autores.

Observa-se na Tabela 1 que as estimativas dos parâmetros são coerentes, com a maior parte dos itens mostrando boa discriminação ($a > 0,7$). A Figura 2 mostra as curvas características de quatro itens. Os itens 22 e 77 discriminam muito bem ($a_{22} = 2,72$ e $a_{77} = 3,17$), sendo o item 77 mais difícil ($b_{22} = -0,38$ e $b_{77} = 1,27$). Por outro lado, o item 62 tem poder de discriminação quase nulo ($a_{62} = 0,17$) e o item 81 possui discriminação moderada $a_{81} = 1,10$.

A classificação dos avaliados na primeira fase do vestibular da Unesp é feita pela teoria clássica (número de acertos), mas se fosse feita pela TRI, o item 62 poderia ser retirado sem grandes prejuízos, já que esse item praticamente não separa avaliados com alto θ de avaliados com baixo θ.

FIGURA 2 – Curvas representando as probabilidades de acerto de quatro itens, em função de θ



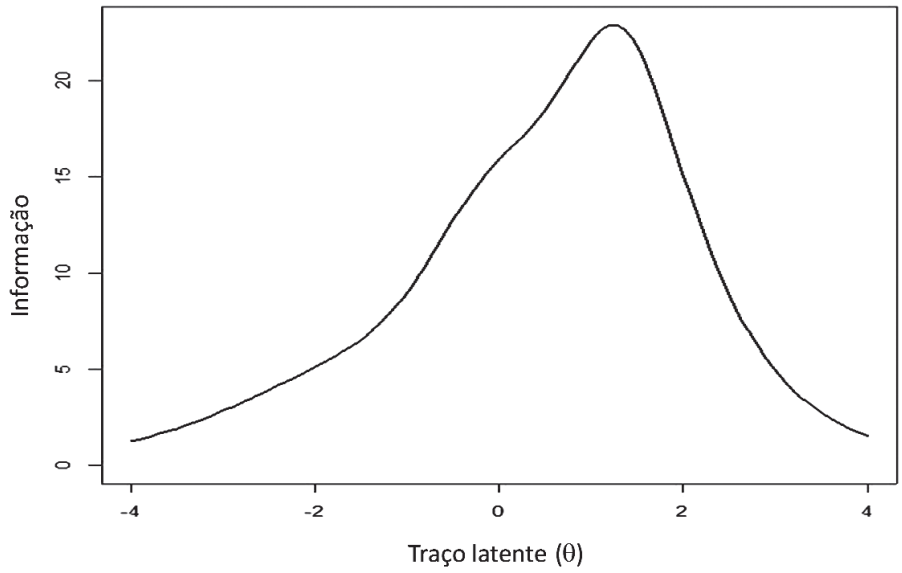
Fonte: Elaboração dos autores.

Observa-se, na Tabela 1, que os itens 21 a 30 (conteúdo de inglês) e 75 a 90 (física e matemática) têm, em geral, grande poder de discriminação (valores altos de *a*). Ou seja, sob o enfoque da TRI são itens que pesariam mais na classificação dos avaliados, especialmente para aqueles com seus θ próximos dos parâmetros de dificuldade, *b*, desses itens. Como a correlação entre o traço latente da TRI e o número de acertos (teoria clássica) é muito forte, pode-se dizer que esses itens têm grande peso na classificação dos avaliados, mesmo que a classificação seja feita pelo número de acertos, e não pela TRI.

Outra característica da TRI é a possibilidade de traçar a curva de informação da prova. Verifica-se, na Figura 3,

que a curva é mais alta na região com θ entre zero e dois, ou seja, esse instrumento fornece mais informação (ou avalia de forma mais precisa) indivíduos com proficiência entre a média e dois desvios padrões acima da média. Essa característica é desejável, pois o maior desafio dos elaboradores dessa prova é classificar adequadamente para a segunda etapa os candidatos de cursos concorridos.

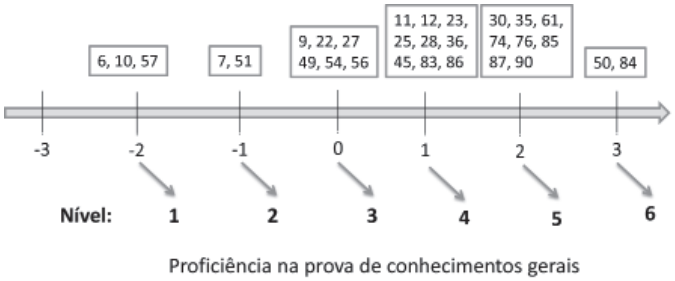
FIGURA 3 – Curva de informação da prova



Fonte: Elaboração dos autores.

Usando as funções de probabilidade dos itens, dadas pela TRI, é possível posicionar os itens na escala do traço latente θ e fazer uma interpretação pedagógica da escala. Como foi utilizada uma escala com distribuição normal de média zero e variância unitária, os itens com maior discriminação foram fixados em cada unidade da escala, basicamente no intervalo de -2 a +3. A Figura 4 apresenta a posição de itens que caracterizam cada ponto da escala, ou seja, itens com alta discriminação e que têm probabilidade superior a 0,60 de serem respondidos por um avaliado com proficiência igual ou maior do que o ponto em que o item está sendo posicionado, muitas vezes chamados na literatura de itens âncoras ou quase-âncoras.

FIGURA 4 – Posicionamento de itens na escala do traço latente θ



Fonte: Elaboração dos autores.

Em Azevedo *et al.* (2013) foi realizada uma interpretação pedagógica dessa escala de seis pontos, a qual é resumida no Quadro 1.

QUADRO 1 – Interpretação dos níveis da escala

NÍVEL 1. O avaliado é capaz de:
Localizar informação apresentada em notícias e fragmentos de texto literário (não ficção).

NÍVEL 2. E mais:
Selecionar informação explícita apresentada em fragmentos de texto literário (não ficção), linguagem gráfica e documentos públicos.
Interpretar informação apresentada em textos de diferentes gêneros, ilustrações, códigos ou mapas.
Estabelecer relações entre imagens e um corpo do texto, escrito em português e/ou inglês, comparando informações pressupostas ou subentendidas.

NÍVEL 3. E mais:
Identificar o sentido de palavras e expressões apresentadas em textos literários (não ficção).
Estabelecer relações entre imagens e um corpo do texto científico para obter dados pressupostos ou subentendidos.
Analisar informações explícitas apresentadas em textos e quadros científicos de média complexidade.
Analisar textos de gêneros distintos para inferir informação.

NÍVEL 4. E mais:
Selecionar informação em textos literários utilizando critérios pré-estabelecidos.
Elaborar proposta com base em informação explícita apresentada em textos, ilustrações e diagramas.
Identificar características específicas associadas a contextos históricos, culturais e tecnológicos.
Interpretar mapas, diagramas para resolver problemas envolvendo cálculos simples.
Relacionar informações para resolver problema utilizando cálculos com operações, funções e relações trigonométricas.

NÍVEL 5. E mais:
Analisar textos de gêneros distintos, apresentados em inglês, para inferir informação.
Comparar diferentes interpretações sobre situações associadas a contextos históricos, sociais, avaliando a validade dos argumentos utilizados.
Analisar texto técnico e científico inferindo e organizando informação subentendida ou pressuposta.
Analisar informações apresentadas em textos técnicos, diagramas e gráficos relacionando-as à determinação de suas características específicas, apresentadas em textos e gráficos.
Resolver problemas envolvendo cálculo de volume de figuras tridimensionais.

NÍVEL 6. E mais:
Relacionar informações apresentadas em textos técnicos complexos para identificar terminologia científica específica.
Resolver problema envolvendo análise combinatória.

Fonte: Azevedo *et al.* (2013).

DIMENSIONALIDADE DA PROVA

Na seção anterior, foi aplicado um modelo unidimensional de TRI numa prova multidisciplinar e interdisciplinar na qual o traço latente pode ser interpretado como uma proficiência geral do avaliado. Contudo, mais importante é verificar quais habilidades e proficiências a prova está realmente medindo. Nesse contexto, primeiro verificam-se quantas dimensões são necessárias para representar bem os itens e os avaliados e, depois, que habilidades e proficiências podem ser identificadas nesse espaço multidimensional.

A abordagem usada foi a análise fatorial de informação completa, baseada no ajuste de modelos TRIM com diferentes quantidades de traços latentes (diferentes dimensões), fazendo comparações sequenciais (2 fatores x 1 fator; 3 fatores x 2 fatores; e assim por diante). Realizando essa análise, verificou-se que a qualidade do ajuste dos modelos foi melhorando até três traços latentes, mas a partir daí foram constatados problemas nesses ajustes. Outras análises, como a análise de componentes principais e a análise paralela sobre a matriz de correlação tetracórica também sugeriram três dimensões.

Do ajuste do modelo com três traços latentes, foram obtidas as cargas fatoriais entre cada item e o traço latente (ou fator, usando a terminologia de análise fatorial). Quanto maior a carga fatorial, maior a relação do item com o correspondente fator. Dessa forma, pode-se interpretar um fator (ou traço latente) considerando o conjunto de itens com alta carga sobre esse fator. A Tabela 2 apresenta as cargas fatoriais destacando aquelas superiores a 0,50.

TABELA 2 – Cargas fatoriais após rotação *oblimin* num modelo de três fatores. Prova de conhecimentos gerais do vestibular da Unesp, 2011

Item	F_1	F_2	F_3	Item	F_1	F_2	F_3	Item	F_1	F_2	F_3
1	-0,06	**0,56**	0,03	**16**	-0,18	**0,61**	-0,05	**31**	0,15	0,15	0,10
2	0,04	0,26	0,04	**17**	-0,08	**0,56**	0,05	**32**	-0,17	0,27	0,05
3	0,41	0,11	0,13	**18**	0,26	0,13	0,22	**33**	0,14	0,41	-0,01
4	0,03	0,39	0,17	**19**	-0,26	**0,73**	0,04	**34**	0,31	0,31	0,02
5	0,06	0,26	0,06	**20**	0,09	0,30	0,12	**35**	**0,51**	0,21	-0,09
6	-0,24	**0,87**	0,03	**21**	-0,05	0,14	**0,85**	**36**	0,26	**0,52**	-0,04
7	-0,29	**0,97**	0,00	**22**	0,10	0,09	**0,79**	**37**	0,24	0,18	-0,02
8	-0,17	**0,74**	-0,02	**23**	-0,03	0,06	**0,69**	**38**	0,13	0,48	-0,04
9	0,07	**0,50**	0,16	**24**	0,08	0,13	**0,77**	**39**	0,47	0,25	-0,05
10	-0,31	**0,84**	0,03	**25**	0,01	0,10	**0,81**	**40**	0,10	**0,60**	-0,03
11	0,29	0,28	0,15	**26**	0,00	0,24	**0,45**	**41**	0,25	0,25	0,01
12	0,21	0,23	0,22	**27**	-0,01	0,17	**0,53**	**42**	-0,01	0,16	-0,02
13	-0,11	0,46	0,01	**28**	0,13	0,05	**0,72**	**43**	0,14	0,27	-0,06
14	-0,12	**0,57**	0,07	**29**	0,15	-0,07	**0,61**	**44**	0,24	0,37	-0,10
15	-0,23	**0,69**	-0,01	**30**	0,07	0,09	**0,61**	**45**	**0,59**	-0,02	0,17

Fonte: Elaboração dos autores. (continua)

TABELA 2 – Cargas fatoriais após rotação *oblimin* num modelo de três fatores. Prova de conhecimentos gerais do vestibular da Unesp, 2011 (continuação)

Item	F_1	F_2	F_3	Item	F_1	F_2	F_3	Item	F_1	F_2	F_3
46	**0,55**	0,20	-0,15	**61**	0,64	0,35	-0,21	**76**	**0,92**	-0,11	0,01
47	0,29	0,37	-0,03	**62**	0,07	0,14	-0,15	**77**	**0,93**	-0,08	-0,02
48	0,30	0,17	-0,01	**63**	0,28	0,14	0,03	**78**	**0,73**	-0,06	0,01
49	**0,75**	-0,01	0,05	**64**	**0,51**	0,09	0,10	**79**	**0,68**	-0,03	0,04
50	não calibrado			**65**	**0,51**	0,07	0,13	**80**	**0,64**	-0,02	0,02
51	0,36	0,30	-0,02	**66**	0,38	0,28	0,06	**81**	0,47	0,05	0,03
52	0,37	0,38	-0,12	**67**	0,34	0,26	0,05	**82**	**0,96**	-0,12	-0,02
53	0,40	0,32	-0,13	**68**	0,12	0,25	-0,02	**83**	**0,95**	-0,17	0,01
54	0,17	0,36	0,07	**69**	0,31	0,12	-0,03	**84**	não calibrado		
55	-0,03	0,49	0,03	**70**	0,48	0,13	0,01	**85**	**0,60**	-0,02	0,14
56	0,05	**0,64**	0,05	**71**	**0,78**	0,12	-0,08	**86**	**0,91**	-0,25	0,10
57	-0,08	**0,57**	-0,03	**72**	**0,64**	0,08	0,01	**87**	**0,88**	-0,16	0,11
58	0,15	**0,55**	0,02	**73**	0,48	0,17	-0,03	**88**	**0,69**	-0,13	0,06
59	0,09	**0,62**	-0,05	**74**	**0,79**	0,02	-0,06	**89**	**0,88**	-0,19	0,01
60	0,09	**0,58**	0,02	**75**	**0,70**	0,03	-0,09	**90**	**0,94**	-0,16	0,01

Fonte: Elaboração dos autores.

Conforme a Tabela 2 e com os itens da prova, verifica-se que o fator 1 está mais associado aos itens de matemática e física (raciocínio lógico); o fator dois aos itens de compreensão

de texto; e o fator três aos itens de língua inglesa. Cabe observar que o fator 1 representa 19,3% da variância das respostas dos avaliados; o fator 2 representa 13,0%; e o fator 3 representa 6%. Além disso, a correlação estimada entre os fatores 1 e 2 é de 0,79; entre 1 e 3 é de 0,67; e entre 2 e 3 é de 0,79. Essas correlações altas entre os fatores podem ser uma das explicações do bom ajuste de um modelo unidimensional, ou seja, de um modelo com um único traço latente.

Outro ponto a se observar é que, exceto os itens associados à língua inglesa, a análise dos padrões de resposta dos avaliados não mostrou uma separação dos itens por disciplina, mas por tipo de traço latente dos avaliados: raciocínio lógico – incluindo conhecimentos específicos – e compreensão de texto.

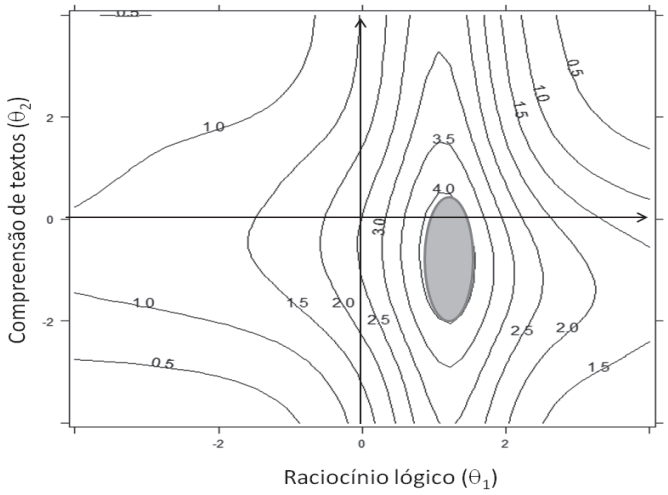
ANÁLISE DA PROVA COM MODELO MULTIDIMENSIONAL

Na seção anterior, identificou-se a presença de três fatores subjacentes nas respostas da prova: raciocínio lógico (F_1), compreensão de texto (F_2) e proficiência em inglês (F_3). Este último pareceu muito bem definido pelos 10 itens dessa disciplina – ver na Tabela 2 que os itens 21 a 30 (itens de língua inglesa) têm cargas fatoriais altas somente em F_3, e nenhum outro item tem carga alta nesse fator. No entanto, há itens com carga elevada apenas em raciocínio lógico (F_1) e em compreensão de texto (F_2), como também itens com cargas moderadas em ambos. Para entender melhor a relação dos itens com os fatores F_1 e F_2, decidiu-se considerar a prova sem os itens de língua inglesa, ajustando um modelo TRI bidimensional.

Para que os eixos cartesianos identificassem melhor os traços latentes raciocínio lógico (F_1) e compreensão de texto (F_2), foram identificados, com o apoio de especialistas da Vunesp, itens tipicamente de compreensão de texto (itens 1, 4, 6, 7, 8, 10, 14, 15, 16, 17, 55, 56, 57 e 59) e itens bem caracterizados de raciocínio lógico (itens 75 a 80 e 82 a 90). No ajuste do modelo, esses itens "puros" foram introduzidos com coeficientes nulos num dos traços latentes. Num passo seguinte, itens com baixo poder discriminatório foram excluídos, resultando numa prova de 66 itens.

A Figura 5 mostra as regiões de maior e de menor informação do teste de 66 itens. Os dois traços latentes considerados têm correlação de 0,67 (moderada positiva). Observa-se que os maiores níveis estão numa região elíptica, identificada com sombreado no gráfico. Assim, pode-se dizer que itens de compreensão de texto discriminam mais bem avaliados com proficiência baixa (abaixo da média), enquanto itens de raciocínio lógico discriminam mais bem avaliados com proficiência alta (acima da média). Além disso, a variação do nível de informação é mais acentuada na horizontal (direção do eixo de raciocínio lógico), mostrando que esses itens só discriminam bem indivíduos de 0,5 a 2 desvios padrões acima da média.

FIGURA 5 – Curvas de níveis da superfície de informação numa prova de 66 itens



Fonte: Elaboração dos autores.

A Tabela 3 mostra as estimativas dos parâmetros de discriminação e de dificuldade dos itens. Verifica-se que os itens associados ao raciocínio lógico (75 a 80 e 82 a 90) têm parâmetros de dificuldade com sinal positivo (itens relativamente difíceis), enquanto os associados à compreensão de texto (1, 4, 6, 7, 8, 10, 14, 15, 16, 17, 55, 56, 57 e 59) têm, em geral, parâmetros de dificuldade com sinal negativo (itens relativamente fáceis), o que se mostra compatível com a análise feita nas curvas de nível da Figura 5.

Assim, se o objetivo fosse o de avaliar separadamente esses dois traços latentes, a prova não seria adequada.

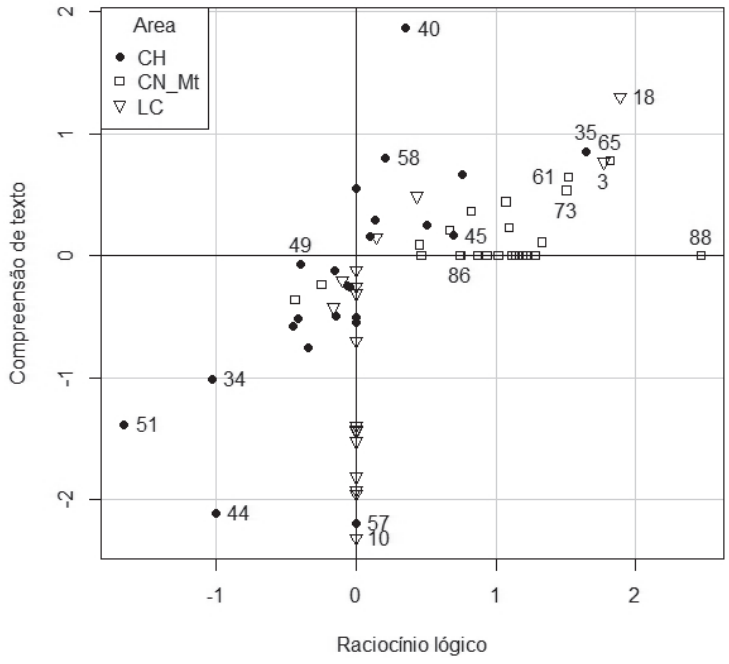
TABELA 3 – Estimativas dos parâmetros de discriminação (a_1 e a_2) e de dificuldade (B) de um modelo TRIM bidimensional em um teste de 66 itens

Item	a_1	a_2	B	Item	a_1	a_2	B	Item	a_1	a_2	B	Item	a_1	a_2	B	Item	a_1	a_2	B
1	0	1,18	-1,81	**17**	0	1,20	-0,25	**46**	1,03	0,51	0,57	**64**	1,20	0,50	1,16	**80**	1,65	0	0,75
3	1,09	0,47	1,93	**18**	0,77	0,53	2,29	**47**	0,58	0,92	0,18	**65**	1,95	0,83	1,98	**81**	0,95	0,30	0,70
4	0	1,21	-0,12	**19**	0,00	1,23	-1,39	**49**	2,12	0,35	-0,41	**66**	0,66	0,54	-0,56	**82**	3,18	0	1,02
6	0	1,74	-1,92	**20**	0,35	0,74	-0,22	**51**	0,83	0,69	-2,15	**67**	0,79	0,76	-0,34	**83**	2,44	0	0,47
7	0	1,98	-1,52	**33**	0,29	1,00	-0,51	**52**	0,68	0,88	-0,73	**70**	0,94	0,42	0,90	**85**	1,79	0	1,28
8	0	1,32	-1,44	**34**	0,76	0,74	-1,45	**53**	0,77	0,67	1,01	**71**	2,41	0,51	1,12	**86**	2,21	0	0,74
9	0,50	1,30	-0,45	**35**	1,12	0,58	1,85	**54**	0,44	0,96	-0,83	**72**	0,91	0,18	0,46	**87**	2,68	0	1,22
10	0,00	1,42	-2,31	**36**	0,66	1,44	0,32	**55**	0	1,18	-0,50	**73**	1,12	0,40	1,60	**88**	0,89	0	2,46
11	0,85	0,95	0,65	**38**	0,29	1,07	-0,25	**56**	0	2,14	-0,55	**74**	2,18	0,18	1,33	**89**	1,98	0	1,19
12	0,73	0,75	0,21	**39**	0,96	0,72	-0,20	**57**	0	1,03	-2,20	**75**	1,61	0	1,21	**90**	2,89	0	1,28
13	0,00	0,78	-1,95	**40**	0,18	0,92	1,89	**58**	0,46	1,75	0,83	**76**	3,11	0	0,94				
14	0	1,16	-0,70	**41**	0,53	0,66	-0,66	**59**	0	1,73	0,55	**77**	3,18	0	1,11				
15	0	1,01	-1,43	**44**	0,38	0,79	-2,33	**60**	0,28	1,57	-0,26	**78**	1,85	0	0,87				
16	0	0,84	-0,31	**45**	1,48	0,35	0,71	**61**	1,62	0,69	1,65	**79**	1,75	0	1,14				

Fonte: Elaboração dos autores.

No ajuste do modelo unidimensional foi possível posicionar os itens numa escala unidimensional (Figura 4). De forma análoga, no ajuste do modelo bidimensional, os itens foram posicionados no plano formado pelos dois traços latentes (Figura 6).

FIGURA 6 – Posicionamento dos itens no plano dos dois traços latentes



Fonte: Elaboração dos autores.

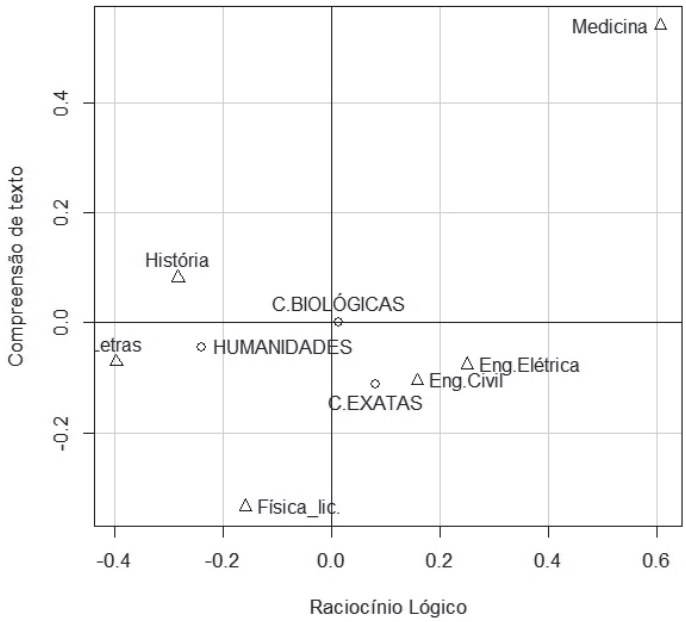
Analisando a Figura 6, verifica-se que os itens de ciências humanas (CH) e linguagens e códigos (LC) posicionam-se com maior variação na vertical (eixo identificado como compreensão de texto), enquanto os itens de ciências da natureza e matemática (CN_Mt) esparramam-se mais na horizontal (eixo do raciocínio lógico), sendo que a maioria dos itens de CN_Mt se posiciona no quadrante superior direito (região dos itens mais difíceis).

Particularizando para alguns itens, verifica-se que os itens de geografia 45 e 49 estão posicionados bastante próximos ao eixo de raciocínio lógico, sendo o 45 com sinal positivo (item mais difícil) e o 49 com sinal negativo (item mais fácil). Lendo o enunciado desses itens em Fundação Vunesp (2011), verifica-se que o item 45 requer o cálculo de fuso horário e o item 49, para ser acertado, envolve o cálculo de distância num mapa, ou seja, procedimentos associados ao raciocínio lógico.

Chama atenção a localização, no quadrante superior direito, de um conjunto de itens que guardam proximidades muito parecidas, tanto em relação ao eixo de raciocínio lógico, quanto ao eixo da compreensão de texto. São itens de língua portuguesa (3), humanidades (35), biologia (61 e 65) e química (73). A leitura desses itens leva ao reconhecimento dos traços comuns que albergam: em todos eles, além da compreensão de textos longos e nem sempre simples, a solução requer a mobilização de conhecimentos específicos da área à qual se associam. É justamente esse grau de proximidade no posicionamento de itens que permite validar a prova no contexto para o qual foi concebida.

A Figura 7 mostra algumas escolhas dos avaliados no mesmo espaço dos itens. Observa-se que, em média, estudantes que optaram pela área de humanidades posicionam-se na direção de itens mais fáceis (quadrante inferior esquerdo). Já a média das proficiências dos estudantes que optaram por ciências exatas afasta-se de humanidades, basicamente na vertical, ou seja, esses estudantes têm, em média, raciocínio lógico mais aguçado.

FIGURA 7 – Posicionamento das médias obtidas pelos avaliados inscritos, segundo a área e segundo alguns cursos



Fonte: Elaboração dos autores.

Analisando a posição média dos estudantes inscritos em alguns cursos (Figura 7), verifica-se que aqueles que optaram por Medicina estão bem acima da média dos demais estudantes, tanto em compreensão de texto como em raciocínio lógico. Os que optaram por engenharias (aqui posicionada apenas a Engenharia Civil e Elétrica), têm proficiência em compreensão de texto ligeiramente abaixo da média, mas raciocínio lógico acima da média. Os que optaram por Letras ou História estão bem abaixo da média em raciocínio lógico, já os que optaram por licenciatura em Física também estão, em média, abaixo da média dos demais estudantes nos dois traços latentes considerados, mas próximo da média quando se trata de raciocínio lógico.

CONSIDERAÇÕES FINAIS

Formalmente, a teoria de resposta ao item (TRI) é usualmente aplicada em instrumentos que se supõe ter apenas um traço latente, ou seja, um instrumento unidimensional. Porém,

muitas vezes é possível aplicar a TRI em instrumentos multidimensionais, desde que se tenha um traço latente que possa representar múltiplas habilidades dos indivíduos avaliados.

Numa prova multidisciplinar, como a prova de conhecimentos gerais da Unesp, espera-se avaliar múltiplas habilidades ou proficiência dos candidatos, mas foi mostrado neste artigo que o modelo de TRI unidimensional ajustou-se bem às respostas dos candidatos, resultando numa escala interpretável pedagogicamente. Todavia, com o ajuste de modelo multidimensional é possível encontrar os diferentes traços latentes presentes na prova e nos avaliados. Nesse caso, verificou-se que era possível representar razoavelmente bem os itens e os avaliados em três eixos cartesianos, denominados de raciocínio lógico, compreensão de texto e proficiência em inglês. Ou seja, embora a prova tenha itens de várias disciplinas, sua dimensionalidade é três.

Em geral, num processo seletivo se quer resumir as habilidades de um indivíduo por um único número, mas a análise de itens num espaço multidimensional permite avaliar melhor as características presentes nesses itens e possibilita verificar quais traços latentes esses itens estão efetivamente medindo. Da mesma forma, com o posicionamento dos indivíduos num espaço multidimensional, torna-se possível identificar suas habilidades predominantes ou então verificar em que habilidades cada indivíduo precisa melhorar. Pensando num processo seletivo de uma universidade, pode-se também verificar em que cursos as habilidades de um candidato seriam mais bem aproveitadas.

A Tabela 4 apresenta dois candidatos hipotéticos, João e Maria, sendo que ambos acertaram 40 itens num total de 66, ou seja, pela teoria clássica, ambos teriam a mesma pontuação. Mas Maria acertou mais itens de compreensão de texto, enquanto João acertou mais itens que exigiam raciocínio lógico. Na escala de média 0 e variância 1, resultante do ajuste do modelo unidimensional da TRI, João ficou posicionado no valor 0,32 e Maria no valor 0,47. Essa pequena diferença deve-se ao fato de que na TRI considera-se o padrão das respostas, incluindo a probabilidade do acerto casual, e não somente o número de acertos. Como os itens de compreensão

de texto eram, em geral, mais fáceis do que os de raciocínio lógico, o padrão de respostas de Maria foi mais coerente com a ideia de que o conhecimento é cumulativo, ou seja, é mais natural acertar os itens fáceis do que os difíceis; no entanto, o padrão de respostas do João não é esperado por um modelo que considera o conhecimento cumulativo e uma única proficiência dominante. Ele acertou itens difíceis e errou muitos itens fáceis. Em outras palavras, a TRI é mais fidedigna na estimação da proficiência dos avaliados.

TABELA 4 – Desempenho de dois candidatos hipotéticos, segundo a metodologia de avaliar a proficiência

	ACERTOS		TRI	TRIM[1]	
Candidatos	quantidade	%	θ	θ_1	θ_2
João	40	66,7	0,32	1,12	0,51
Maria	40	66,7	0,47	-0,14	1,63

[1] θ_1 = Raciocínio lógico; θ_2 = Compreensão de texto.
Fonte: Elaboração dos autores.

Com a abordagem da TRIM, pode-se verificar que os dois candidatos são bem diferentes: enquanto João tem boa habilidade em raciocínio lógico, Maria tem boa habilidade em compreensão de texto. Esse tipo de análise permite orientar os candidatos para os cursos mais compatíveis com as suas habilidades.

O posicionamento de itens e indivíduos num espaço multidimensional ainda é assunto incipiente na área de avaliação educacional. Têm surgido muitas propostas de modelagem e de análise com a TRIM, mas ainda não se dispõe de uma metodologia consagrada como na teoria clássica ou na TRI unidimensional. Este artigo procurou fazer alguns ensaios tomando como estudo de caso o vestibular da Unesp.

AGRADECIMENTOS

À Fundação para o Vestibular da Unesp – Vunesp, pela cessão dos resultados do Vestibular Unesp 2011 – primeira fase; e pelo incentivo dado no desenvolvimento de pesquisas em avaliação educacional. Agradecemos, também, ao Prof. Dalton F. de Andrade (UFSC e Vunesp) pelas suas valiosas sugestões.

REFERÊNCIAS

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C. *Teoria da Resposta ao Item*: conceitos e aplicações. São Paulo: Associação Brasileira de Estatística, 2000.

AYALA, R. J. *The theory and practice of Item Response Theory*. New York: The Guilford, 2009.

AZEVEDO, T. C. A. M; TREVISAN, L. M. V.; ROCHA, G. T.; BARBETTA, P. A.; ANDRADE, D. F.; TAVARES, H. Uma escala de proficiência baseada na descrição de conhecimentos e habilidades dos candidatos em um processo seletivo para ingresso na universidade. In: REUNIÃO DA ASSOCIAÇÃO BRASILEIRA DE AVALIAÇÃO EDUCACIONAL, 7., 2013, Brasília, DF. *Anais*... Brasília, DF.: ABAVE, 2013.

BARBETTA, P. A.; ANDRADE, D. F.; BORGATTO, A. F. Análise de provas do Enem segundo modelos de TRI multidimensionais. In: CONGRESSO BRASILEIRO DE TEORIA DA RESPOSTA AO ITEM, 2., 2011, Salvador, BA. *Anais*... Salvador: CONBRATRI, 2011.

BOCK, R. D.; AITKIN, M. Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, v. 46, n. 4, p. 443-459, 1981.

BOCK, R. D.; GIBBONS, R.; MURAKI, E. Full-Information Item Factor Analysis. *Applied Psychological Measurement*, v. 12, n. 3, p. 261-280, 1988.

CHALMERS, P. *Mirt*: A multidimensional Item Response Theory. Package for the R Environment. *Journal of Statistical Software*, v. 48, n. 6, p. 1-29, 2012. Disponível em: <www.jstatsoft.org/v48/i06/>. Acesso em 30 ago. 2013.

CHILDS, R. A.; OPPLER, S. H. Implication of test dimensionality for unidimensional IRT scoring: an investigation of a High-Stake Testing Program. *Education and Psychological Measurement*, v. 60, p. 939-955, 2000.

DIAZ, A. M. M. Multidimensional Item Response Theory Models where the ability has a latent linear structure. 2010. Tese (Doutorado) – Universidad Nacional de Colombia, Colômbia. 2010.

FUNDAÇÃO CESGRANRIO. *SAEB*. Relatórios técnicos: análise clássica do teste e análise da Teoria da Resposta ao Item. Rio de Janeiro: Fundação CESGRANRIO, 2004.

FUNDAÇÃO VUNESP. *Relatório vestibular UNESP 2011*. Disponível em: <http://www.vunesp.com.br/pesquisaUnesp/relatorios/2011.pdf>. Acesso em: 10 fev. 2013.

LAROS, J. A.; PASQUALI, L.; RODRIGUES, M. M. M. *Análise da unidimensionalidade das provas do SAEB*. RelatórioTécnico do CPAE – UnB, 2000.

PRIMI, R.; SILVA, M. C. R.; SANTANA, P. R.; MUNIZ, M.; ALMEIDA, L. S. The use of the bi-factor model to test the uni-dimensionality of a battery of reasoning tests. *Psicothema*, Oviedo, v. 25, p. 115-122, 2013.

RECKASE, M. *Multidimensional Item Response Theory*. USA: Springer, 2009.

VITÓRIA, F.; ALMEIDA, L. S.; PRIMI, R. Unidimensionalidade em testes psicológicos: conceito, estratégias e dificuldades na sua avaliação. *PSIC – Revista de Psicologia da Vetor Editora*, v. 7, n. 1, p. 1-7, 2006.

WIRTH, R. J.; EDWARDS, M. C. Item factor analysis: current approaches and future directions. *Psychological Methods*, n. 12, p. 58–79, 2007.

PEDRO A. BARBETTA
Professor associado do Departamento de Informática e Estatística e do Programa de Mestrado Profissional Métodos e Gestão em Avaliação da Universidade Federal de Santa Catarina (UFSC). Pesquisador da Fundação Vunesp
pedro.barbetta@ufsc.br

LIGIA M. V. TREVISAN
Professora e assessora de Diretoria da Fundação Vunesp
ligiamvtrevisan@gmail.com

HELITON TAVARES
Professor associado da Faculdade de Estatística e do Programa de Pós-Graduação em Matemática e Estatística da Universidade Federal do Pará (UFPA). Pesquisador da Fundação Vunesp
heliton@globo.com

TÂNIA C. A. DE MACEDO AZEVEDO
Professora da Faculdade de Engenharia da Universidade Estadual Paulista "Júlio de Mesquita Filho" (Unesp), Campus de Guaratinguetá. Superintendente acadêmico da Fundação Vunesp
tmacedo@feg.unesp.br

Recebido em: AGOSTO 2013
Aprovado para publicação em: FEVEREIRO 2014