http://dx.doi.org/10.18222/eae.v30i74.5324

DIMENSIONALIDADE E ESCALA DE PROFICIÊNCIA EM UMA PROVA INTERDISCIPLINAR

LIGIA MARIA VETTORATO TREVISAN[I]
PEDRO ALBERTO BARBETTA[II]
DALTON FRANCISCO DE ANDRADE[III]
GUARACY TADEU ROCHA[IV]
TÂNIA CRISTINA ARANTES DE MACEDO AZEVEDO[V]

RESUMO

O artigo apresenta uma análise da dimensionalidade da prova de conhecimentos gerais do vestibular da Universidade Estadual Paulista "Júlio de Mesquita Filho" (Unesp) e consolida a construção de uma escala em uma prova que inclui itens associados às diferentes disciplinas integrantes do currículo do ensino médio no Estado de São Paulo. Embora a prova seja interdisciplinar, esse estudo mostra a viabilidade de se adotar uma medida unidimensional pela teoria da resposta ao item. Além disso, por meio de uma análise fatorial de informação completa, foi possível identificar quais habilidades e competências a prova está medindo. Os fatores que mais se destacaram foram o raciocínio lógico, a proficiência em Língua Inglesa e o conhecimento em Humanidades.

PALAVRAS-CHAVE TEORIA DA RESPOSTA AO ITEM • ESCALA DE AVALIAÇÃO • ANÁLISE FATORIAL • VESTIBULAR.

I Fundação para o Vestibular da Universidade Estadual Paulista (Vunesp), São Paulo-SP, Brasil; http://orcid.org/0000-0003-2506-9656; *ligiamvtrevisan@gmail.com*

II Universidade Federal de Santa Catarina (UFSC), Florianópolis-SC, Brasil; http://orcid.org/0000-0002-5359-0134; *pedro.barbetta@ufsc.br*

III Universidade Federal de Santa Catarina (UFSC), Florianópolis-SC, Brasil; http://orcid.org/0000-0002-4403-980X; *dalton.andrade@ufsc.br*

IV Universidade Estadual Paulista "Júlio de Mesquita Filho" (Unesp), Botucatu-SP, Brasil; http://orcid.org/0000-0002-6538-2762; *grocha@vunesp.com.br*

V Universidade Estadual Paulista "Júlio de Mesquita Filho" (Unesp), Guaratinguetá-SP, Brasil; http://orcid.org/0000-0002-9841-3086; *tcristinamacedo@gmail.com*

DIMENSIONALIDAD Y ESCALA DE CONOCIMIENTO EN UNA PRUEBA INTERDISCIPLINARIA

RESUMEN

En este trabajo se presenta un análisis de la dimensionalidad de la prueba de conocimientos generales del examen de ingreso a la Universidad Estadual Paulista "Júlio de Mesquita Filho" (Unesp) y se consolida la construcción de una escala en un test que incluye ítems asociados a las diferentes disciplinas que forman parte del currículo de la educación secundaria en el estado de São Paulo. Aunque la prueba es interdisciplinaria, este estudio demuestra la viabilidad de adoptar una medida unidimensional por la Teoría de Respuesta al Ítem. Además, por medio de un análisis factorial de información completa fue posible conocer las habilidades y competencias que la prueba mide. Los factores que más se destacaron fueron el razonamiento lógico, el dominio del idioma inglés y el conocimiento de Humanidades.

PALABRAS CLAVE TEORÍA DE LA RESPUESTA AL ÍTEM • ESCALA DE EVALUACIÓN • ANÁLISIS FACTORIAL • EXAMEN DE INGRESO A LA UNIVERSIDAD.

DIMENSIONALITY AND PROFICIENCY SCALE AT AN INTERDISCIPLINARY TEST

ABSTRACT

This paper analyzes the dimensionality of the general knowledge test for admission into the Universidade Estadual Paulista "Júlio de Mesquita Filho" (Unesp) and consolidates the construction of a scale in a test that includes items associated with the various subjects forming the high school curriculum in the state of São Paulo. Although the test is an interdisciplinary one, this study shows the feasibility of adopting a one-dimensional measure through item response theory. In addition, we used full information factor analysis to map the skills and competencies measured by the test. The main factors found were logical reasoning, proficiency in English and knowledge of Humanities.

KEYWORDS ITEM RESPONSE THEORY • ASSESSMENT SCALE • FACTOR ANALYSIS • ADMISSION EXAM.

INTRODUÇÃO

Um grande desafio para selecionar avaliados num vestibular é responder à questão: até que ponto a nota de um avaliado está refletindo suas habilidades e seus conhecimentos, especialmente numa prova interdisciplinar que engloba todo o conteúdo do ensino básico?

Muitas avaliações de larga escala têm uma matriz de referência por área do conhecimento e interpretações pedagógicas da escala de pontuação, como é o caso do Saeb (Sistema de Avaliação do Ensino Básico). Nele, tanto em Língua Portuguesa quanto em Matemática, é possível verificar, em cada faixa da escala de pontuação, quais habilidades o estudante provavelmente domina. Essa escala é apresentada para o 5º e 9º anos do ensino fundamental e para a 3ª série do ensino médio (BRASIL, 2015).

Para construir uma escala desse tipo, é necessário, primeiramente, verificar se é razoável resumir as habilidades e conhecimentos do avaliado para um único número – sua nota geral. O próprio Enem (Exame Nacional do Ensino Médio) tem, atualmente, uma matriz de referência para cada uma das quatro áreas avaliadas por provas objetivas (BRASIL, 2015), mas, para classificar os candidatos, as universidades normalmente usam a média aritmética de cinco notas: quatro de provas objetivas e uma de redação.

Contudo, há apenas alguns trabalhos acadêmicos que verificam as propriedades dessa medida resultante, como os de Vieira (2016) e Gomes (2018).

A consistência interna e a dimensionalidade de provas interdisciplinares aplicadas no Brasil foram abordadas por alguns autores, dentre os quais estão Quaresma (2014), que estudou a prova da primeira fase do vestibular da Fuvest, e Coelho (2014), que analisou a prova do Exame Nacional de Desempenho dos Estudantes (Enade), na área de Estatística, mas nenhum deles tentou realizar uma escala com interpretação pedagógica para essas provas.

Com base na edição de 2011 do vestibular da Universidade Estadual Paulista "Júlio de Mesquita Filho" (Unesp), Barbetta *et al.* (2014) mostraram que os resultados da prova de conhecimentos gerais são bem representados por um traço latente tridimensional, mas também indicaram que era razoável fazer uma redução de dimensionalidade, simplificando para uma única dimensão, ou seja, cada avaliado podia ter apenas uma nota retratando a composição de habilidades que ele domina. Neste trabalho, o estudo é ampliado para várias edições da prova (2011-2014), fazendo uma análise mais elaborada da dimensionalidade e completando a interpretação da escala.

Como bem pondera Pasquali (2003), a dimensionalidade deve ser considerada uma questão de grau, uma vez que o desempenho humano é multideterminado e multimotivado. Dessa forma, é razoável se perguntar quão bem a proficiência geral medida pela prova pode representar de forma satisfatória um constructo mais complexo, envolvendo vários fatores subjacentes. Reckase (2009) argumenta que a utilização da teoria da resposta ao item (TRI) em dados multidimensionais pode gerar uma medida resumo de várias habilidades de um indivíduo.

O presente trabalho visa, de certa forma, avaliar a validade das provas do vestibular da Unesp. A validade de um teste (prova) corresponde ao grau em que a evidência e a teoria apoiam as interpretações dos escores produzidos pelo teste, considerando os seus propósitos (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, 2014, p. 11). Evidências de conteúdo estão presentes na prova de conhecimentos gerais da Unesp, uma vez que os itens são elaborados e avaliados em conformidade com as diretrizes curriculares nacionais para o ensino médio (BRASIL, 2013), os parâmetros curriculares nacionais para o ensino médio (BRASIL, 2000) e o currículo do estado de São Paulo (SÃO PAULO, 2011a, 2011b, 2011c, 2011d). Neste trabalho é apresentada uma escala de proficiência interpretada com base no posicionamento de itens da prova e descritores baseados nas habilidades e competências das publicações supracitadas.

A fonte de validade mais presente neste estudo é a estrutura interna da prova. As análises da estrutura interna podem indicar o grau em que as relações entre os itens da prova estão em conformidade com o construto em questão. Essa estrutura é analisada por índices baseados na teoria clássica dos testes (TCT), na análise da dimensionalidade e na interpretação de fatores evidenciados por análise fatorial.

O artigo contempla uma apresentação da prova de conhecimentos gerais do vestibular da Unesp, a descrição dos métodos adotados, a análise das provas pela TCT e pela TRI, uma proposta de interpretação pedagógica da medida gerada pela TRI e a análise da dimensionalidade da prova de 2014, incluindo análise fatorial baseada em modelos da TRI uni e multidimensionais. Com essa última análise, buscam-se fatores subjacentes às respostas da prova associados com o que a prova está medindo.

A PROVA DE CONHECIMENTOS GERAIS DO VESTIBULAR DA UNESP

O presente trabalho baseia-se na análise dos resultados das provas de conhecimentos gerais dos vestibulares da Unesp de 2011 a 2014. A prova de conhecimentos gerais pode ser considerada inter e multidisciplinar, sendo composta de 90 questões (itens) de múltipla escolha, organizadas nas diferentes áreas especificadas nos parâmetros curriculares nacionais do ensino médio.

Os colegiados superiores da Unesp têm definido, sistematicamente, entre os objetivos de seus concursos vestibulares, selecionar candidatos capazes de: articular ideias de modo coerente; compreender ideias, relacionando-as; expressar-se com clareza; e conhecer o conteúdo do currículo da educação básica do estado de São Paulo. Esses requisitos são importantes na medida em que sinalizam um processo seletivo guiado pela investigação de aspectos cognitivos da formação adquirida pelos estudantes em seu percurso na educação básica. Assim, tanto pelos conteúdos tratados quanto pela forma como é proposta esta abordagem aos candidatos em cada uma das áreas examinadas, a preparação da prova de conhecimentos gerais atende às orientações que para ela são estabelecidas. O resultado dessa preparação é um instrumento de avaliação em que os itens de 1 a 30 são de linguagens, códigos e suas tecnologias (Língua Portuguesa, Literatura, Língua Inglesa, Educação Física e Arte), os itens de 31 a 60 são de Ciências Humanas e suas tecnologias (História, Geografia e Filosofia) e os itens de 61 a 90 são de Ciências da Natureza, Matemática e suas tecnologias (Biologia, Química, Física e Matemática).

A correção das provas é realizada pela TCT e a nota do avaliado é proporcional ao seu número de acertos.

Os candidatos mais bem classificados na prova de conhecimentos gerais, conforme o curso de graduação pretendido, são selecionados para a etapa seguinte, composta por uma prova de conhecimentos específicos, com questões dissertativas e uma redação. A classificação final considera o desempenho do candidato em todas as etapas da seleção: provas de conhecimentos gerais e conhecimentos específicos e redação. Cabe observar que atualmente o candidato pode optar por levar em conta a nota do Enem na composição da nota de conhecimentos gerais. Além disso, há reserva de vagas para estudantes que cursaram integralmente o ensino médio em escolas públicas.

Os dados das respostas, sem identificação dos candidatos, foram gentilmente cedidos pela Fundação para o Vestibular da Universidade Estadual Paulista (Vunesp), entidade responsável pela elaboração, aplicação e correção das provas. As questões das provas, gabaritos e algumas estatísticas estão disponíveis no *site* da Vunesp.[1]

MÉTODOS

Como a Vunesp corrige as provas pela teoria clássica dos testes, a análise inicial foi feita sob essa abordagem.

Na TCT, uma preocupação básica é que a prova tenha alta consistência interna, ou seja, que todo item possua correlação positiva forte ou moderada com o número total de acertos. Para avaliar a consistência interna da prova, foi calculado o chamado coeficiente α de Cronbach. Para verificar a qualidade do item no contexto de consistência interna, avaliou-se a variação desse coeficiente com a retirada do item em análise, além da correlação bisserial (r_{bis}) entre o item e o total de acertos calculado com os demais itens da prova. A base teórica dessas medidas é descrita em Revelle (2017a, cap. 7) e o cálculo foi feito com o pacote psych (REVELLE, 2017b) do *software* R (R CORE TEAM, 2017).

Para os propósitos deste artigo, realizou-se uma análise mais elaborada, a partir da teoria da resposta ao item (TRI). Os modelos de TRI relacionam a probabilidade de um avaliado acertar um item com parâmetros desse item e com a proficiência do avaliado. Existem vários modelos de TRI tratados na literatura, conforme ampla descrição em Van der Linden (2016). O presente

1 Relatório Vestibular Unesp. Disponível em: http://www.vunesp.com.br/Institucional/EstatisticaVestibular.

trabalho adota modelo similar ao usado no Saeb e no Enem. No Saeb é utilizada a função *probit* de três parâmetros; neste trabalho fez-se uso do mesmo modelo empregado no Enem – o logístico de três parâmetros –, cuja probabilidade de um indivíduo com proficiência *j* acertar o item *i* é dada por:

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

Onde os parâmetros *a*, *b* e *c* são relativos ao item e o parâmetro θ é associado ao avaliado. Mais especificamente:

a_i representa o nível de discriminação do item *i*;

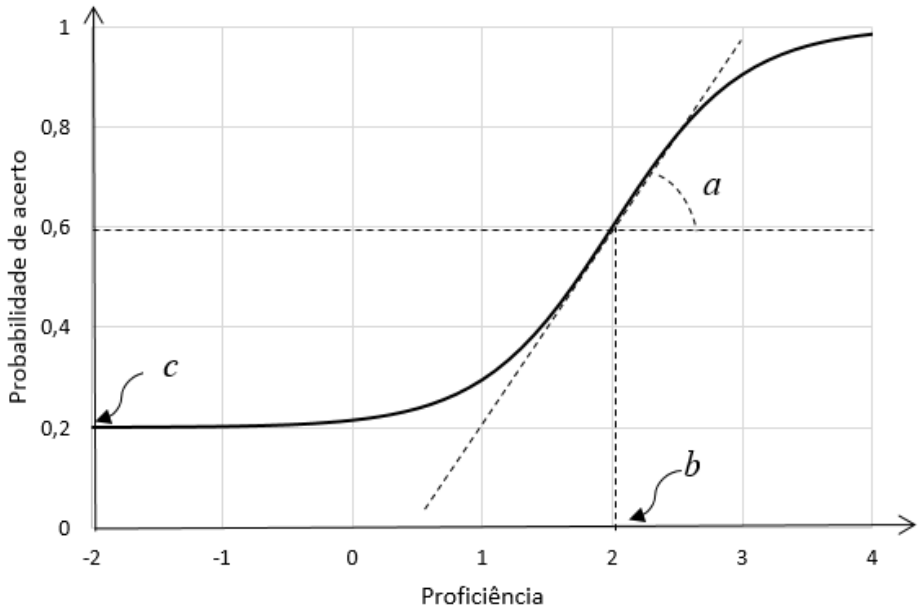
b_i corresponde ao nível de dificuldade do item *i*;

c_i refere-se à probabilidade de acerto casual do item *i*;

θ_j representa o traço latente do avaliado *j*, supostamente com distribuição normal de média 0 e desvio padrão 1.

O parâmetro de dificuldade *b* está na mesma escala de proficiência θ dos avaliados. Pela suposição de θ ter distribuição normal padrão, seus valores usualmente estão entre -3 e +3. Assim, um item com $b = 2$ pode ser considerado difícil para um avaliado com proficiência mediana ($\theta = 0$), mas pode ser visto como fácil para um avaliado com proficiência $\theta = 3$ (Gráfico 1). O parâmetro de discriminação *a* indica quão bem o item discrimina entre avaliados de proficiência abaixo de *b* e avaliados acima de *b*, sendo desejável $a > 1$. Visualizando no gráfico, o parâmetro *a* é tão maior quanto mais inclinada for a curva no ponto *b*.

GRÁFICO 1 – Curva da probabilidade de acerto de um item com *a* = 2, *b* = 2 e *c* = 0,2, em função da proficiência θ



Fonte: Elaboração dos autores, 2018.

A estimação dos parâmetros do modelo foi realizada com o pacote mirt (CHALMERS, 2012, 2017) do *software* livre R (R CORE TEAM, 2017), adotando o método da máxima verossimilhança marginal.

A avaliação da qualidade dos ajustes dos modelos foi feita pelas estatísticas RMSEA (erro quadrático médio de aproximação), CFI (índice de ajuste comparativo) e TLI (índice de Tucker-Lewis), usando a função M2 do mirt. Conforme argumentam Thimoty (2015) e Cai e Hansen (2013), essas estatísticas têm a vantagem de ser pouco sensíveis ao tamanho da amostra, ao contrário do usual teste qui-quadrado de comparação de modelos, que, usualmente, rejeita a adequação de modelos quando a amostra é muito grande, mesmo nos casos em que os modelos são adequados.

O posicionamento dos itens na escala de proficiência foi realizado com os chamados itens âncora conforme descrito em Andrade, Tavares e Valle (2000). Esse posicionamento constituiu a base para a interpretação pedagógica da escala.

Numa etapa posterior, realizou-se um estudo de dimensionalidade da prova. Um procedimento usual para avaliação da dimensionalidade de um instrumento (prova) é a análise de componentes principais baseada na matriz de correlações formada entre os pares dos itens. No presente caso, foram adotadas as chamadas correlações tetracóricas, próprias para itens dicotômicos. Um

complemento da análise de componentes principais é a análise paralela, feita com a simulação de amostras aleatórias com itens não correlacionados, permitindo uma espécie de teste não paramétrico para a avaliação da dimensionalidade do instrumento. Essas técnicas são descritas por vários autores, em particular Revelle (2017a), Olsson, Drasgow e Dorans (1982) e Garrido, Abad e Ponsoda (2013).

Uma análise mais completa da dimensionalidade foi realizada pela chamada análise fatorial de informação completa, que é baseada em modelos de teoria da resposta ao item multidimensional (TRIM), ou seja, trata-se de uma extensão do modelo de TRI, na qual o traço latente θ é considerado multidimensional, composto por vários fatores. Neste trabalho adotou-se a chamada família de modelos de TRIM compensatórios, conforme descrito por Reckase (2009), que é a formulação mais usual de modelos multidimensionais. Em termos computacionais, novamente foi usado pacote mirt (CHALMERS, 2012, 2017).

ANÁLISE DAS PROVAS PELA TCT

A Tabela 1 apresenta, para cada edição da prova, o coeficiente α, a média dos coeficientes de correlação bisserial (r_{bis}), o número de itens que contribuem negativamente para a consistência interna (reduzem α) e o número de itens com r_{bis} negativo ou muito baixo. Essas estatísticas foram obtidas pelo pacote computacional psych (REVELLE, 2017b).

TABELA 1 – Estatísticas clássicas de consistência interna das provas

EDIÇÃO	NÚMERO DE AVALIADOS	COEF. α	r_{bis} MÉDIO	NÚMERO DE ITENS	
				α É REDUZIDO COM A RETIRADA	$r_{bis} < 0,15$
2011	73.178	0,919	0,32	7	7
2012	82.840	0,915	0,31	12	10
2013	84.393	0,926	0,34	4	4
2014	88.739	0,919	0,32	13	10

Fonte: Elaboração dos autores, 2018.

A maioria dos autores considera o instrumento de medida com consistência interna satisfatória quando $\alpha > 0,70$. Nas quatro edições examinadas, o coeficiente α de Cronbach foi superior a 0,90, indicando alta consistência interna.

Na análise dos itens da prova, verifica-se que, dos 360 itens que compõem as quatro edições, apenas 36 reduzem α quando retirados, e 31 apresentam

coeficiente bisserial muito pequeno (r_{bis} < 0,15). A título de comparação, adotando o mesmo procedimento numa amostra aleatória de 25 mil concluintes do ensino regular, para a prova de Matemática do Enem de 2013, foi obtido α = 0,89 e r_{bis} médio de 0,35, ou seja, valores muito próximos dos encontrados no presente estudo.

Em síntese, as estatísticas da TCT mostram que as provas de conhecimentos gerais da Unesp, nas quatro edições examinadas, têm forte evidência de validade por consistência interna, mesmo considerando suas características inter e multidisciplinares.

ANÁLISE DAS PROVAS VIA TRI UNIDIMENSIONAL

Num primeiro momento, analisou-se separadamente cada edição do vestibular. Poucos itens tiveram seus parâmetros estimados com valores ruins (mal calibrados), tais como coeficiente de discriminação abaixo de 0,5, parâmetro de dificuldade fora do intervalo [-5; 5] ou erros padrão relativamente altos. Os itens 32, 42 e 62 da edição de 2011, os itens 2 e 31 da edição de 2012 e os itens 43, 52, 69 e 87 da edição de 2014 ficaram mal calibrados. Esses nove itens estão incluídos naqueles em que o coeficiente de correlação bisserial foi muito baixo, conforme quantitativo apresentado na Tabela 1.

A Tabela 2 mostra algumas estatísticas de adequação do ajuste de modelos TRI, obtidas com auxílio do pacote computacional mirt. Nesta análise foram excluídos os nove itens que apresentaram problemas de calibração.

TABELA 2 – Estatísticas de qualidade do ajuste dos modelos de TRI

EDIÇÃO	RMSEA	TLI	CFI
2011	0,017	0,983	0,984
2012	0,014	0,988	0,989
2013	0,016	0,987	0,988
2014	0,018	0,982	0,983

Fonte: Elaboração dos autores, 2018.

Os metodologistas consideram o ajuste adequado quando RMSEA < 0,05 e TLI e CFI são maiores que 0,90 (THIMOTY, 2015, p. 74). Assim, pelos resultados da Tabela 2, os modelos de TRI estão muito bem ajustados.

Uma possível crítica em se adotar um modelo unidimensional, quando se espera traço latente multidimensional, é que a dependência entre itens pode

não ser totalmente explicada pela diferença de proficiências entre os avaliados, contrariando um pressuposto básico da TRI: a independência local, a qual pode ocorrer por não se considerar a dimensão adequada do instrumento.

Seguindo a abordagem de Chen e Thissen (1997), a análise de possível dependência local pode ser feita por meio dos resíduos da matriz de correlações entre itens, após ajuste de um modelo da TRI. Eventuais correlações moderadas entre esses resíduos apontariam dependência local. Segundo Schilling (2009), uma heurística para avaliar os resíduos é verificar se a raiz quadrada da média das correlações residuais quadráticas é menor que 0,05 e se existem poucas correlações superiores a 0,10, condições que sugerem um instrumento essencialmente unidimensional.

No presente estudo essa análise foi realizada com auxílio do pacote computacional mirt na edição de 2014. A raiz quadrada da média das correlações residuais quadráticas foi de 0,02. Apenas cinco correlações, em módulo, foram maiores que 0,10, dentre as 3.655 correlações calculadas. Esses resultados sugerem que a prova pode ser considerada essencialmente unidimensional.

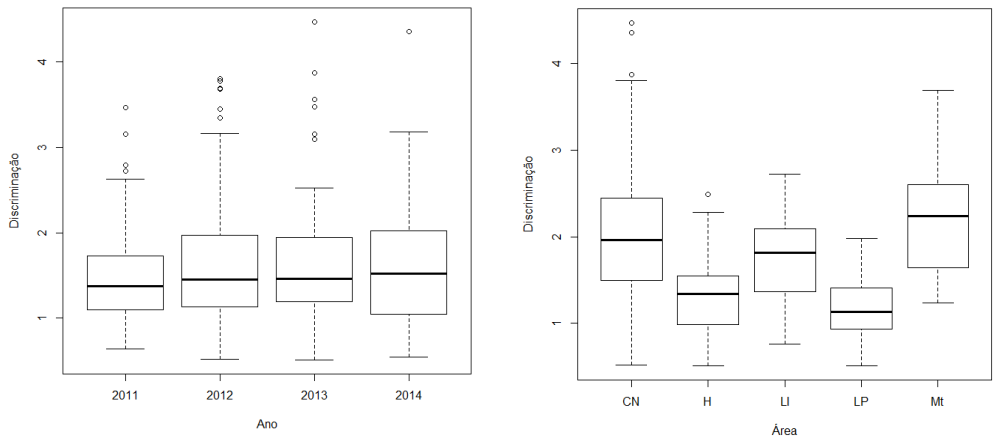
Uma grande vantagem da TRI em relação à TCT é a possibilidade de várias provas serem colocadas numa mesma escala, permitindo comparabilidade entre elas. Esse processo é conhecido na literatura como equalização e, no presente trabalho, verificou-se a possibilidade de fazer a equalização via população (ANDRADE; TAVARES; VALLE, 2000, p. 81). Geralmente esse processo é feito com um conjunto de avaliados realizando as várias provas, mas também é possível quando as populações puderem ser supostas equivalentes (KOLEN; BRENNAN, 2004, p. 298).

A questão de supor as populações equivalentes é um tanto subjetiva. Neste estudo, tal suposição foi feita levando em conta que o Índice de Desenvolvimento da Educação Básica (Ideb) do ensino médio do estado de São Paulo manteve-se praticamente constante, com valores de 4,1 em 2011, 4,1 em 2013 e 4,2 em 2015, em escolas públicas e privadas. Cabe ressaltar, também, que nesse período não houve mudança nos critérios e objetivos do processo seletivo, além de serem mantidos os referenciais curriculares que orientam a seleção dos conhecimentos requisitados pelas provas aos candidatos. Também não ocorreu, no período, mudança significativa dos candidatos quanto ao perfil socioeconômico, tipo de escola na qual cursaram o ensino fundamental e o ensino médio (pública ou privada), frequência a cursos preparatórios para vestibular, período transcorrido entre a conclusão do ensino médio e o vestibular, escolaridade dos pais, entre outras variáveis obtidas a partir das respostas dos candidatos ao questionário socioeconômico preenchido por ocasião da inscrição para o vestibular do respectivo ano.

Considerando o exposto, ao fazer a calibração conjunta das quatro edições, têm-se as proficiências praticamente na mesma escala, possibilitando ampliar a abrangência histórica do estudo. Essa calibração foi feita com 351 itens, correspondentes aos 90 itens de cada edição, após a retirada dos nove que apresentaram problemas de calibração na análise isolada de cada prova.

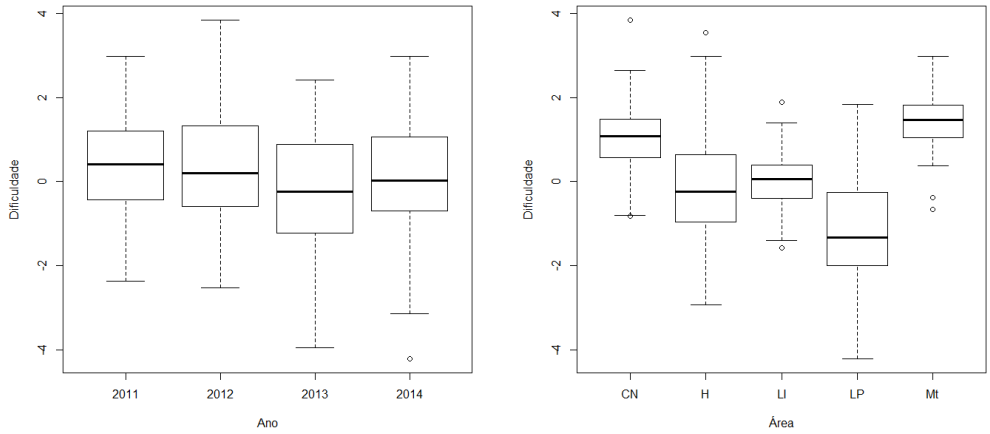
Os gráficos 2 a 4 apresentam diagramas de caixas (*boxplots*) das estimativas dos três parâmetros do modelo de TRI por ano e área de conhecimento: Humanidades (H), Língua Portuguesa e Literatura (LP), Língua Inglesa (LI), Ciências da Natureza (CN) e Matemática (Mt). Os itens em inglês foram colocados em separado, porque foi verificado em estudos anteriores que eles eram posicionados em dimensão diferente dos itens de Língua Portuguesa e Literatura (BARBETTA *et al.*, 2014). Também foram separados os itens de Matemática e Ciências da Natureza. Os gráficos foram elaborados com funções do *software* R.

GRÁFICO 2 – Estimativas dos parâmetros de discriminação (*a*), por ano e área de conhecimento



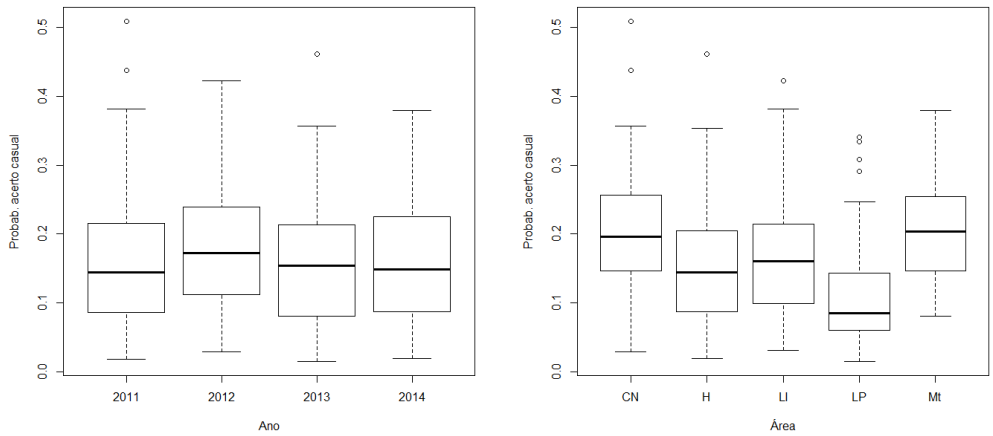
Fonte: Elaboração dos autores, 2018.

GRÁFICO 3 – Estimativas dos parâmetros de dificuldade (*b*), por ano e área de conhecimento



Fonte: Elaboração dos autores, 2018.

GRÁFICO 4 – Estimativas dos parâmetros de acerto casual (*c*), por ano e área de conhecimento



Fonte: Elaboração dos autores, 2018.

Verifica-se pouca diferença da distribuição dos parâmetros da TRI entre as quatro edições do vestibular, valendo apenas ressaltar que a prova de 2013 apresenta leve tendência de itens mais fáceis (Gráfico 3). Já em termos das áreas do conhecimento, observa-se que os itens de Ciências da Natureza (CN) e Matemática (Mt) tendem a ser mais difíceis e discriminam mais. Já os itens de Humanidades (H) e Língua Portuguesa e Literatura (LP) são, em geral, mais fáceis e discriminam menos.

Para efeito de comparação entre a proficiência obtida pela TRI nas quatro edições agregadas e o escore calculado pela TCT em cada edição, calculou-se o coeficiente de correlação de Pearson entre as duas medidas, obtendo o valor de 0,967, indicativo de correlação bastante forte, o que mostra a congruência dos dois processos de mensuração. Isso não quer dizer que as duas abordagens levem a resultados similares para o propósito dessa prova, pois, na classificação de candidatos, pequenas diferenças na pontuação podem resultar em posições bem diferentes em um mesmo curso. Além disso, a TRI gera resultados mais justos por considerar a coerência do padrão de respostas do avaliado e, ainda, permite interpretar a escala de pontuação e, sob certas condições, se ter a comparabilidade das notas em várias edições.

CONSTRUÇÃO E INTERPRETAÇÃO DA ESCALA UNIDIMENSIONAL

Esta seção apresenta uma escala unidimensional para a prova de conhecimentos gerais da Unesp, levando em conta que o principal objetivo da prova é classificar candidatos para uma segunda etapa, na qual as provas são dissertativas. Medidas unidimensionais em provas interdisciplinares já foram propostas em outros trabalhos, como em Quaresma (2014), na prova da primeira fase da Fuvest, Coelho (2014), no Enade, área de Estatística, e Vieira (2016), no Enem, considerando as quatro provas objetivas como uma prova única.

Não é comum, porém, fazer interpretação pedagógica de uma medida unidimensional em uma prova interdisciplinar. No Enem, por exemplo, constrói-se uma escala interpretável para cada área, considerando que essas áreas têm matrizes de referência próprias, as quais são compostas de várias habilidades específicas da área. Contudo, quando são utilizadas as provas do Enem para classificar avaliados nas universidades, trabalha-se com uma medida resumo: em geral, a média aritmética das provas das quatro áreas e da redação. Essa medida síntese carece de interpretação pedagógica.

Seguindo Reckase (2009, p. 184), a análise por modelos unidimensionais de TRI pode resultar numa medida de referência composta de uma prova que

avalia múltiplas habilidades. Nesse sentido, o presente artigo faz um ensaio ao interpretar pedagogicamente uma prova interdisciplinar, adotando termos gerais que possam englobar os conteúdos de várias disciplinas.

Com a aplicação da TRI unidimensional, os candidatos dos vestibulares da Unesp podem ser posicionados numa escala contínua, em que a origem representa um candidato de proficiência média, e cada unidade da escala corresponde a um desvio padrão que se afasta da média, a escala (0, 1). O interessante na metodologia da TRI é que os itens também podem ser posicionados na mesma escala de proficiência dos avaliados. Neste artigo adota-se a abordagem descrita em Andrade, Tavares e Valle (2000, p. 110) para o posicionamento dos itens na escala.

A escala foi interpretada em seis níveis, sendo o nível 1 estabelecido em dois desvios padrão abaixo da média e o nível 6 em três desvios padrão acima da média. Essa assimetria deve-se à natureza da prova, destinada a discriminar melhor candidatos acima da média. O Quadro 1 mostra o posicionamento dos itens considerados âncora ou quase âncora, ou seja, itens que estão fortemente associados a um dado nível da escala. Os critérios para definição dos itens âncora foram baseados em Andrade, Tavares e Valle (2000, p. 110) e Beaton e Allen (1992), que consideram as seguintes restrições: probabilidade de acerto no nível posicionado maior que 0,65; probabilidade de acerto no nível anterior menor que 0,50; e diferença entre o nível posicionado e o nível anterior maior que 0,30. Para a categorização em "quase âncora" não foi incluída a última restrição, mas exigiu-se que o nível de discriminação do item fosse maior que 1.

QUADRO 1 – Composição de itens na escala, identificando o ano de aplicação e o conteúdo do item

	ÂNCORA				QUASE ÂNCORA			
NÍVEIS	**2011**	**2012**	**2013**	**2014**	**2011**	**2012**	**2013**	**2014**
Nível 1	10-Lp		3-Lp	6-Lp	6-Lp		6-Lp	
			7-Lp				48-Geo	
			10-Lp		57-Fil			
Nível 2	7-Lp	6-Lp	9-Lp	10-Lp	14-Lp	14-Lp	14-Lp	12-Lp
	51-Geo	18-Lp	16-Lp	51-Geo	21-Ing		45-Geo	47-Geo
		58-Fil	18-Lp				52-Geo	
		12-Lp	22-Ing		34-His			
			25-Ing		44-His			
Nível 3	9-Lp	22-Ing	20-Lp	23-Ing	7-Lp	7-Lp	47-Geo	33-His
	49-Geo	23-Ing	21-Ing	28-Ing	27-Ing	46-Geo	68-Bio	
	22-Ing	26-Ing	23-Ing	33-His	46-Geo	60-Fil	88-Mat	
	56-Fil	59-Fil	49-Geo	37-His	54-Geo			
		60-Fil	58-Fil	38-His	60-Fil			
			60-Fil	40-His	67-Bio			
			61-Bio	45-Geo				
				49-Geo				
Nível 4	11-Lp	24-Ing	26-Ing	21-Ing	23-Ing	47-Geo	13-Lp	67-Bio
	25-Ing	25-Ing	41-His	68-Bio	59-Fil		28-Ing	2-Lp
	36-His	65-Bio	42-His	72-Quim			30-Ing	
	83-Fis	29-Ing	62-Bio	73-Quim			31-His	
	45-Geo	66-Bio	64-Bio	74-Quim			35-His	
	12-Lp	70-Quim	69-Quim	88-Mat			67-Bio	
	86-Mat	68-Quim	71-Quim				77-Fis	
	28-Ing	71-Quim	73-Quim					
			82-Fis					
Nível 5	3-Lp	40-His	70-Quim			58-Fil		44-Geo
	35-His	64-Bio	75-Quim					88-Mat
	61-Bio	67-Bio	76-Fis					
	74-Quim	73-Quim	78-Fis					
	90-Mat	77-Fis	79-Fis					
	30-Ing	81-Fis	81-Fis					
	79-Mat	83-Mat	83-Fis					
	76-Fis	85-Mat	84-Fis					
	77-Fis	90-Mat	90-Mat					
	85-Mat							
	87-Mat							
Nível 6	50-Geo	38-His	38-His	65-Bio				
	84-Mat	84-Mat	74-Quim	66-Bio				
	50-Geo	89-Mat	86-Mat	79-Fis				
				80-Fis				

Fonte: Elaboração dos autores, 2018.

Conteúdos: Lp=Língua Portuguesa e Literatura; Geo=Geografia; Fil=Filosofia; Ing=Inglês; His=História; Mat=Matemática; Bio=Biologia; Quim=Química; Fis=Física.

No que se refere à distribuição dos conteúdos associados aos itens em cada nível da escala, o quadro permite verificar que, ao longo das edições da prova, os itens de cada matéria estão distribuídos de forma parecida e posicionados em vários níveis de proficiência. Itens de Geografia e História estão distribuídos por todos os níveis. O mesmo ocorre, em menor intensidade, com as questões de Língua Inglesa e Língua Portuguesa. Já os itens de Ciências da Natureza e suas Tecnologias concentram-se entre os níveis 3 e 4 da escala, mas fica mais evidente que, nas disciplinas que integram essa área, eles são mais presentes nos níveis mais elevados, sugerindo maior complexidade das habilidades a eles associadas.

Em relação aos estudos anteriores, o resultado do tratamento conjunto das quatro provas mostrou que quase todos os itens âncora de 2012 a 2014 puderam ser incluídos na escala que já existia (ver BARBETTA *et al.*, 2014). No entanto, para que a descrição ficasse mais pertinente, foi necessário adequar os termos de alguns descritores, para incluir novos tipos de texto como estímulo na composição dos contextos apresentados ao candidato – por exemplo, textos científicos, textos não literários e gráficos.

A utilização do recurso da calibração conjunta dos itens de quatro edições do vestibular da Unesp e a suposição da semelhança do público respondente levaram à estimativa dos parâmetros dos itens numa mesma escala, na qual os itens foram posicionados para a interpretação pedagógica.

Cabe observar que as provas do vestibular da Unesp baseiam-se nas diretrizes curriculares do Estado de São Paulo (SÃO PAULO, 2011a, 2011b, 2011c, 2011d), em que são apresentados os conteúdos e habilidades esperadas pelos estudantes em cada área do conhecimento e cada etapa do ensino. A interpretação da escala de proficiência geral foi baseada no posicionamento dos itens e levou em conta as diretrizes curriculares, procurando descrever habilidades e conhecimentos em linguagem que pode ser facilmente compreendida nas várias áreas que compõem a prova de conhecimentos gerais.

O Quadro 2 apresenta uma proposta de interpretação pedagógica dos níveis da escala, observando que, em cada nível, o aluno deve ter as habilidades descritas nesse nível e nos níveis anteriores (propriedade cumulativa do conhecimento).

QUADRO 2 – Interpretação dos níveis da escala de proficiência geral

Nível 1	• Localizar informação apresentada em notícias e fragmentos de texto literário (não ficção).
Nível 2	• 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	• Identificar o sentido de palavras e expressões, apresentadas em textos literários (não ficção). • Identificar os principais elementos dos sistemas políticos, econômicos e culturais de organização da vida social. • 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, inclusive em língua inglesa. • Analisar textos de gêneros distintos para inferir informação.
Nível 4	• Selecionar informação em textos literários utilizando critérios preestabelecidos. • Elaborar proposta com base em informação explícita apresentada em textos, ilustrações e diagramas, inclusive em língua inglesa. • Identificar características específicas associadas a contextos históricos, culturais, científicos e tecnológicos. • Interpretar mapas, diagramas para resolver problemas envolvendo cálculos simples. • Relacionar informações para resolver problemas utilizando cálculos com operações, funções e relações trigonométricas. • Analisar informações apresentadas em partes, em textos científicos de média complexidade, e entender a inter-relação existente entre as partes. • Aplicar conhecimento científico específico na resolução de problemas.
Nível 5	• Analisar textos de gêneros distintos, apresentados em inglês, para inferir informação. • Comparar diferentes interpretações sobre situações associadas a contextos históricos e 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, gráficos ou figuras. • Resolver problemas envolvendo cálculo de volume de figuras tridimensionais. • Analisar gráficos de alta complexidade para inferir informação técnica específica.
Nível 6	• Relacionar informações apresentadas em textos técnicos complexos, nas diferentes áreas do conhecimento, para identificar terminologia, fatos ou características próprias da área. • Resolver problema envolvendo análise combinatória. • Aplicar leis e conceitos fundamentais da física para resolver problemas que envolvem movimento.

Fonte: Elaboração dos autores, 2018, complementando o quadro apresentado em Barbetta *et al*. (2014).

Embora o Saeb tenha matriz de referência, por ano escolar e área, as habilidades em cada nível da escala são descritas de forma ligeiramente diferente, geralmente mais detalhadas, já que a descrição considera o conteúdo dos itens posicionados (BRASIL, 2015). Gomes (2018), que fez interpretação pedagógica

das notas globais do Enem – todas as áreas conjuntamente –, encaixou parte das habilidades das matrizes de referência nos níveis da escala. No presente artigo, a análise foi feita em provas inter e multidisciplinares. Dessa forma, algumas habilidades do Quadro 2 foram escritas de forma mais genérica para englobar diferentes áreas do conhecimento.

ESTUDO DA DIMENSIONALIDADE DA PROVA

A prova de conhecimentos gerais da Unesp é inter e multidisciplinar, já que alguns itens perpassam os conteúdos curriculares e demandam para sua resolução a mobilização de conhecimentos associados a diferentes disciplinas. Nesse contexto, surge a questão: é razoável supor que itens e avaliados possam ser bem representados por um traço latente unidimensional? Em outras palavras, é razoável uma única medida, construída com as respostas da prova, diferenciar bem os avaliados?

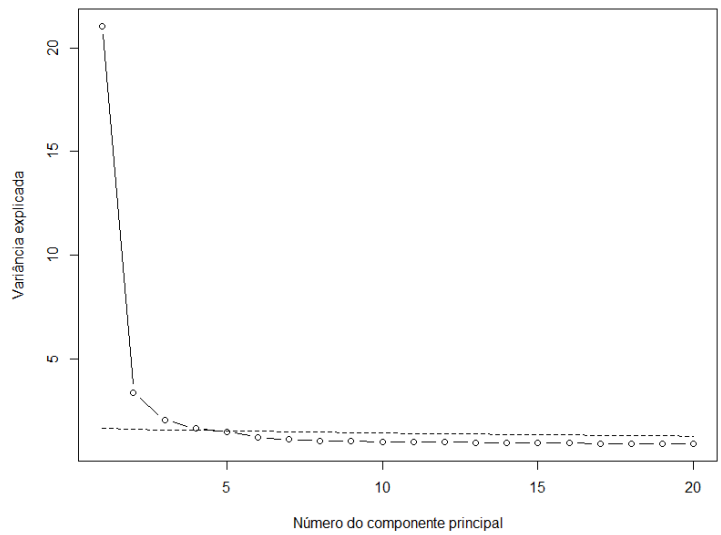
Schilling (2009) diferencia unidimensionalidade restrita de unidimensionalidade essencial, em que o segundo conceito está associado a uma dimensão dominante, suficiente para a aplicação da TRI unidimensional. Conforme comentado anteriormente, Reckase (2009, p. 184) pondera que uma medida produzida pela TRI unidimensional pode representar uma referência composta de várias habilidades. Nesse contexto, a prova de conhecimentos gerais do vestibular da Unesp pode ter um fator dominante, composto de diversas habilidades dos avaliados.

Cabe destacar alguns estudos em provas nacionais. Quaresma (2014) verificou que a prova da primeira fase da Fuvest é bem representada por quatro dimensões, mas um modelo da TRI unidimensional de três parâmetros também se ajustou bem. Vieira (2016) considerou as quatro provas objetivas do Enem, edição 2012, como uma prova única de 4 x 45 = 180 itens, mostrando que havia uma dimensão amplamente dominante. Além disso, a proficiência gerada pela TRI apresentou correlação bastante alta com a média aritmética simples das proficiências das quatro áreas publicadas pelo Inep. Barbetta *et al.* (2014) mostraram que a prova de conhecimentos gerais da Unesp, edição de 2011, é bem representada por um traço latente tridimensional, mas um modelo da TRI unidimensional também é bem ajustado, indicando que o traço latente unidimensional é uma composição de habilidades.

Para realizar a análise de dimensionalidade, é importante obter correlações entre todos os pares de itens, o que não é possível fazer com base nas respostas das quatro edições agregadas. Assim, a análise da dimensionalidade foi feita somente com os avaliados da prova de 2014.

O Gráfico 5 apresenta o resultado da análise de componentes principais sobre a matriz de correlação tetracórica. A linha pontilhada refere-se à análise paralela realizada com amostras simuladas, em que os itens são estatisticamente independentes. Essa análise foi feita com auxílio do pacote computacional "psych" do *software* R.

GRÁFICO 5 – Variâncias explicadas pelas componentes principais



Fonte: Elaboração dos autores, 2018.

Observa-se, no Gráfico 5, que o primeiro componente principal tem variância explicada bastante superior aos demais, fato que evidencia uma dimensão dominante e, assim, permite o uso da TRI unidimensional. Contudo, há outros pontos acima da linha pontilhada, a qual representa o limite tolerável, segundo a chamada análise paralela. Isso sugere que, apesar de ser possível aplicar a TRI unidimensional, um modelo de TRIM deve se ajustar melhor aos padrões de respostas dos avaliados.

Com o objetivo de elucidar as habilidades que estão combinadas no traço latente unidimensional, interpretado como proficiência geral dos avaliados, foram ajustados, com auxílio do pacote computacional mirt do *software* R, vários modelos de TRIM, desde o modelo com um traço latente ou fator (modelo unidimensional) até aquele com sete traços latentes (modelos multidimensionais). A qualidade desses ajustes foi avaliada pelas estatísticas

TLI e CFI, discutidas anteriormente. Também foi verificado o critério de informação da *deviance* (DIC), discutido em Weakliem (1999). Essa última estatística geralmente cresce à medida que se aumenta o número de dimensões do modelo ajustado, mas a elevação é relativamente pequena quando se incluem dimensões supérfluas. A Tabela 3 mostra os resultados.

TABELA 3 – Medidas de qualidade do ajuste em modelos de uma a sete dimensões

DIMENSÃO	DIC	REDUÇÃO DO DIC	TLI	CFI
1	7.882.498	-	0,982	0,983
2	7.830.303	52.195	0,993	0,994
3	7.806.596	23.707	0,996	0,997
4	7.799.396	7.200	0,997	0,997
5	7.794.582	4.814	0,996	0,996
6	7.792.613	1.969	0,912	0,926
7	7.792.175	438	0,952	0,960

Fonte: Elaboração dos autores, 2018.

Analisando as estatísticas da Tabela 3, verifica-se que há forte redução do DIC ao passar do modelo de uma para o de duas dimensões, e do modelo de duas para o de três dimensões. A partir desse estágio, as reduções são mais suaves, especialmente após cinco dimensões. As estatísticas TLI e CFI apresentam maiores valores nos modelos de três, quatro e cinco dimensões. Em suma, a dimensionalidade dessa prova é mais bem representada por três, quatro ou cinco fatores, tanto na observação das estatísticas da Tabela 3 quanto no Gráfico 5.

Em Barbetta *et al.* (2014) foram ajustados modelos com três fatores, os quais foram interpretados como: compreensão de textos e conhecimentos gerais, especialmente evidenciado pelos itens de Humanidades, Língua Portuguesa e Literatura; raciocínio lógico e conhecimentos específicos, bastante evidenciados por itens de Física e Matemática; e proficiência em inglês. Nesse estudo, por restrições computacionais da época, houve algumas limitações nos dados e no processo de estimação dos parâmetros do modelo.

No presente trabalho, usando a totalidade dos dados de 2014 e estimando todos os parâmetros dos itens simultaneamente, o ajuste com três fatores (modelo TRIM de três fatores) apresentou resultado parecido com o trabalho anterior, embora o primeiro fator, denominado anteriormente de compreensão de texto e conhecimentos gerais, tenha ficado agora mais associado aos itens de Humanidades, enquanto os itens de Linguagens e Códigos (exceto de

Língua Inglesa) não ficaram tão bem representados. Por outro lado, a solução com cinco fatores mostrou resultados mais interessantes. A Tabela 4 apresenta as cargas fatoriais do ajuste com cinco fatores, em que é possível inferir o que cada fator está medindo.

TABELA 4 – Cargas fatorais da prova de 2014, via análise fatorial de informação completa, incluindo rotação Oblimin

ITEM[1]	F1	F2	F3	F4	F5	ITEM[1]	F1	F2	F3	F4	F5
1	0,14	**0,33**	0,03	-0,17	**0,36**	46	**0,61**	0,07	0,08	0,05	0,09
2	-0,06	0,15	**0,30**	0,29	0,26	47	**0,48**	-0,03	0,12	0,19	0,10
3	-0,03	0,21	0,07	-0,13	**0,51**	48	**0,60**	0,08	0,09	0,09	-0,03
4	0,12	0,23	0,18	-0,06	**0,44**	49	**0,60**	0,06	0,12	0,02	0,11
5	0,17	0,10	0,10	0,01	**0,34**	50	**0,59**	-0,04	0,05	-0,07	0,24
6	0,25	0,24	0,10	-0,16	**0,37**	51	**0,51**	0,06	0,13	0,01	0,24
7	0,18	0,17	**0,37**	-0,23	0,10	52	-	-	-	-	-
8	0,14	0,07	0,19	0,21	**0,37**	53	0,51	0,15	-0,03	-0,20	0,09
9	**0,34**	-0,04	0,03	0,03	0,13	54	**0,68**	0,08	0,05	0,05	0,08
10	-0,03	0,23	0,22	0,11	**0,42**	55	0,28	-0,06	0,10	0,03	0,28
11	0,19	0,02	0,06	0,02	**0,46**	56	**0,48**	-0,09	-0,07	-0,02	0,25
12	0,25	0,14	-0,01	-0,05	**0,43**	57	**0,45**	0,03	0,19	0,07	0,24
13	0,07	0,16	0,06	0,27	0,24	58	**0,49**	-0,01	0,06	0,11	0,21
14	0,15	-0,01	0,05	0,09	**0,56**	59	**0,64**	-0,05	0,12	-0,04	0,09
15	0,18	0,03	0,05	0,01	**0,42**	60	**0,34**	-0,07	0,10	-0,03	0,27
16	0,23	0,04	0,17	0,11	0,41	61	**0,33**	0,17	0,09	**0,41**	0,06
17	0,27	0,01	0,16	-0,01	**0,33**	62	**0,33**	**0,37**	0,02	0,21	0,06
18	-0,06	0,24	0,13	0,27	0,24	63	**0,34**	0,08	0,07	0,05	0,04
19	-0,06	0,26	**0,33**	0,24	**0,34**	64	**0,39**	**0,33**	0,10	0,09	0,02
20	0,19	-0,07	0,18	0,08	**0,41**	65	0,14	**0,65**	0,17	-0,24	0,15
21	0,09	0,07	**0,70**	0,00	0,02	66	**0,56**	0,20	0,02	**0,37**	-0,16
22	0,05	-0,06	**0,97**	-0,04	-0,02	67	**0,36**	0,20	0,10	0,05	-0,04
23	0,11	-0,01	**0,84**	-0,07	0,05	68	**0,38**	0,24	0,07	0,12	0,06
24	-0,01	0,01	**0,98**	-0,06	-0,07	69	-	-	-	-	-
25	0,12	0,12	**0,71**	-0,10	-0,02	70	0,15	**0,52**	0,09	0,29	0,07
26	-0,05	-0,03	**0,94**	0,09	0,02	71	**0,35**	0,28	0,10	**0,40**	0,00
27	0,01	0,04	**0,48**	0,02	0,07	72	**0,31**	0,25	0,11	**0,41**	-0,03
28	0,03	-0,08	**0,91**	0,00	0,07	73	0,16	**0,34**	0,10	**0,43**	0,05
29	-0,07	0,04	**0,80**	0,18	0,03	74	0,14	**0,35**	0,09	**0,37**	0,07

(*Continua*)

(*Continuação*)

ITEM[1]	F1	F2	F3	F4	F5	ITEM[1]	F1	F2	F3	F4	F5
30	-0,08	0,09	**0,94**	-0,03	-0,03	75	0,21	0,30	0,02	**0,47**	0,04
31	**0,46**	0,13	0,14	0,07	0,02	76	0,15	0,20	0,05	**0,41**	0,11
32	**0,63**	0,07	0,08	-0,07	-0,12	77	0,14	**0,53**	0,04	**0,30**	0,04
33	**0,58**	0,04	0,15	0,01	0,09	78	0,10	**0,67**	0,08	0,18	0,02
34	0,28	0,04	0,07	-0,04	0,26	79	0,23	**0,74**	0,08	0,06	-0,10
35	**0,51**	-0,03	0,09	0,18	0,15	80	0,18	**0,68**	0,12	0,10	-0,13
36	**0,32**	0,03	0,10	-0,05	-0,04	81	-0,01	**0,80**	0,09	0,05	-0,04
37	**0,51**	0,01	0,10	0,02	0,23	82	0,01	**0,63**	0,07	0,06	0,03
38	**0,58**	0,01	0,14	0,05	0,22	83	0,15	**0,61**	0,02	0,10	-0,11
39	**0,46**	0,01	0,10	-0,09	0,24	84	0,00	**0,89**	0,03	-0,09	0,02
40	**0,69**	0,10	0,09	-0,04	0,08	85	0,00	**0,60**	0,11	0,24	0,10
41	**0,46**	0,03	0,08	0,00	0,06	86	0,05	**0,76**	0,02	-0,06	0,08
42	**0,79**	0,06	-0,01	0,14	-0,11	87	-	-	-	-	-
43	-	-	-	-	-	88	-0,02	**0,83**	-0,06	0,06	0,17
44	**0,35**	0,15	0,08	0,24	-0,03	89	-0,14	**0,87**	0,01	-0,01	0,06
45	**0,63**	0,25	0,05	0,04	-0,08	90	0,16	0,22	0,17	0,11	0,24

Fatores	F1	F2	F3	F4	F5
Variância explicada	10,2	8,8	8,2	2,6	3,8

Fonte: Elaboração dos autores, 2018.

(1) Itens de 1 a 30 são de Linguagem e Códigos, de 31 a 60 são de Humanidades e de 61 a 90 são de Ciências da Natureza e Matemática.

A proficiência em Inglês é representada por um fator próprio – são os itens de 21 a 30 que estão bastante correlacionados com o fator F3. A habilidade denominada em trabalho anterior de compreensão de texto e conhecimentos gerais agora é representada pelos fatores F1 e F5, sendo F1 mais associado aos itens de Humanidades (31 a 60) e F5 aos itens de Língua Portuguesa e Literatura (1 a 20).

O fator F2 é bem representado pelos itens de Física e Matemática (77 a 90), sendo assim um fator relacionado com o raciocínio lógico dos avaliados. Já a maioria dos itens de Química (69 a 76) apresenta carga moderada no fator F4 e alguns também têm associação com o fator F2. Os itens de Biologia (61 a 68) têm mais associação com o fator F1, ou seja, estão junto com os itens de Humanidades, mas alguns desses também dividem carga com os fatores F2 e F4 (raciocínio lógico e proficiência em Química, respectivamente). Em suma, a representação da prova em cinco fatores pode ser assim descrita:

- compreensão de texto e conhecimentos gerais na área, com destaque para Humanidades (F1);
- raciocínio lógico (F2);
- proficiência em Inglês (F3);
- proficiência em Química (F4);
- proficiência em Língua Portuguesa e Literatura (F5).

Vale a pena destacar que os fatores com maior variância explicada (última linha da Tabela 4) são F1, F2 e F3. O fator F1, associado principalmente aos itens de Humanidades, e o fator F2 (proficiência em Língua Portuguesa e Literatura) têm correlação positiva moderada. Esses dois fatores também estão associados à habilidade de compreensão de texto.

Ressalta-se que os fatores aqui descritos não constituem por si só medidas dos construtos compreensão de texto, raciocínio lógico etc., mas, sim, medidas compostas pelos itens das provas associados a esses construtos.

Os resultados foram parecidos com os encontrados por Quaresma (2014), que analisou a prova da primeira fase da Fuvest. A análise estatística do autor apontou a adequação entre três e cinco fatores, porém ele enfatizou a solução por quatro fatores. Ele não encontrou relação direta com as quatro grandes áreas adotadas a partir de 2009 no Enem, mas interpretou os quatro fatores que sobressaíram na análise: "habilidade geral", por se relacionar bem com itens de treinamento e de leitura e compreensão; "domínio de raciocínio lógico", associado a itens de treinamento; "domínio de análise interpretativa", que envolve itens de várias disciplinas; e "domínio da língua inglesa".

CONSIDERAÇÕES FINAIS

A prova de conhecimentos gerais da Unesp é inter e multidisciplinar e é corrigida segundo a TCT, sendo a pontuação baseada no número de acertos. Neste trabalho avaliou-se a validade da prova, no contexto de sua estrutura interna, baseando-se, especialmente, na TRI. Mesmo considerando a interdisciplinaridade da prova, verificou-se haver um fator bem dominante, que foi interpretado como o "desempenho geral do avaliado", um traço latente unidimensional em que se fez uma proposta de escala de seis níveis interpretada pedagogicamente.

Em outra vertente, realizou-se uma análise fatorial para identificar os fatores subjacentes que compõem essa prova, e isso se justifica com base no fato de que o conhecimento do que um instrumento de medida está realmente medindo permite usá-lo com maior eficiência. O estudo mostra que

a técnica para geração de um valor único para o avaliado e a análise resultada em múltiplos fatores associados às habilidades desse candidato não são contraditórias. Na verdade, as análises se completam, apesar de no primeiro caso ter-se a ideia de unidimensionalidade do instrumento de medida, enquanto no segundo a ideia é de multidimensionalidade.

Na análise, foram agregadas as provas de 2011 a 2014, obtendo uma medida em escala única. O coeficiente de correlação dessa medida com o número de acertos dos candidatos de cada prova foi de 0,967, correlação bastante forte, mostrando a congruência dos dois processos de mensuração. Ressalta-se, porém, a melhor adequação da medida baseada na TRI para os propósitos deste estudo.

Num estudo complementar, buscando entender empiricamente o que a prova está medindo, verificou-se que um modelo pentadimensional é bem ajustado às respostas dos avaliados. Com a análise deste resultado, foi possível verificar como os itens das diferentes disciplinas interagem.

Enquanto os resultados do ajuste de um modelo unidimensional permitem classificar os avaliados por uma medida geral de desempenho, o modelo multidimensional possibilita conhecer melhor o que a prova está medindo. Em especial, foram identificados três fatores com alto poder explicativo, que estão associados a humanidades, raciocínio lógico e proficiência em inglês. Também merecem registro os itens de Química e os de Língua Portuguesa e Literatura, que aparecem destacados em outros dois fatores.

AGRADECIMENTOS

À Fundação para o Vestibular da Unesp – Vunesp pela cessão de dados do vestibular da Unesp e pelo incentivo à pesquisa.

REFERÊNCIAS

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION; AMERICAN PSYCHOLOGICAL ASSOCIATION; NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION. *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association, 2014.

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

BARBETTA, P. A.; TREVISAN, L. M. V.; TAVARES, H. R.; AZEVEDO, T. C. A. M. Aplicação da teoria da resposta ao item uni e multidimensional na análise da prova de conhecimentos gerais do vestibular da Unesp. *Estudos em Avaliação Educacional*, São Paulo, v. 25, n. 57, p. 280-302, jan./abr. 2014.

BEATON, A. E.; ALLEN, N. L. Interpreting scales through scale anchoring. *Journal of Educational Statistics*, Washington, v. 17, n. 2, p. 191-204, June 1992.

BRASIL. Ministério da Educação. *Parâmetros curriculares nacionais para o ensino médio*. Brasília, 2000. Disponível em: portal.mec.gov.br/expansao-da-rede-federal/195-secretarias-112877938/seb-educacao-basica-2007048997/12598-publicacoes-sp-265002211. Acesso em: 7 dez. 2018.

BRASIL. Ministério da Educação. *Diretrizes curriculares nacionais para educação básica*. Brasília, 2013. Disponível em: http://portal.mec.gov.br/docman/julho-2013-pdf/13677-diretrizes-educacao-basica-2013-pdf/file. Acesso em: 7 dez. 2018.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Matrizes e escalas*. Brasília, 2015. Disponível em: http://provabrasil.inep.gov.br/escalas-de-proficiencia. Acesso em: 26 set. 2018.

CAI, L.; HANSEN, M. Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, Leicester, UK, v. 66, n. 2, p. 245-276, May 2013.

CHALMERS, P. Mirt: a multidimensional item response theory package for the R Environment. *Journal of Statistical Software*, Innsbruck, Austria, v. 48, n. 6, p. 1-29, May 2012. Disponível em: www.jstatsoft.org/v48/i06/. Acesso em: jun. 2015.

CHALMERS, P. *Package mirt*: multidimensional item response theory. Version 1.25. 2017. Disponível em: http://cran.r-project.org/web/packages/mirt/mirt.pdf. Acesso em: set. 2017.

CHEN, W. H.; THISSEN, D. Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, Washington, v. 22, n. 3, p. 265-289, Sept. 1997.

COELHO, E. C. *Teoria da resposta ao item*: desafios e perspectivas em exame multidisciplinar. 2014. 189 f. Tese (Doutorado em Métodos Numéricos em Engenharia) – Universidade Federal do Paraná, Curitiba, 2014.

KOLEN, M. J.; BRENNAN, R. L. *Test equating, scaling, and linking*: methods and practices. New York: Springer, 2004.

GARRIDO, L. E.; ABAD, J. A.; PONSODA, V. A new look at horn's parallel analysis with ordinal variables. *Psychological Methods*, Washington, v. 18, n. 4, p. 454-474, Oct. 2013.

GOMES, D. E. *Avaliação pedagógica para uma escala única do Enem*. 2018. Dissertação (Mestrado em Métodos e Gestão em Avaliação) – Universidade Federal de Santa Catarina, Florianópolis, 2018.

OLSSON, U.; DRASGOW, F. E.; DORANS, N. The polyserial correlation coefficient. *Psychometrika*, Switzerland, v. 47, n. 3, p. 337-347, Sept. 1982.

PASQUALI, L. *Psicometria*: teoria dos testes na psicologia e na educação. Petrópolis: Vozes, 2003.

QUARESMA, E. S. *Modelagem para construção de escalas avaliativas e classificatórias em exames seletivos utilizando teoria da resposta ao item uni e multidimensional*. 2014. 187 f. Tese (Doutorado em Ciências) – Escola Superior de Agricultura "Luiz de Queiroz" (Esalq), Universidade de São Paulo, Piracicaba, 2014.

R CORE TEAM. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2017. Disponível em: www.r-project.org. Acesso em: set. 2017.

RECKASE, M. *Multidimensional item response theory*. New York: Springer, 2009.

REVELLE, W. *An introduction to psychometric theory with applications in R*. FreeTextbookList.com. 2017a. Disponível em: http://personality-project.org/r/book/. Acesso em: 8 set. 2017.

REVELLE, W. *Psych*: procedures for psychological, psychometric, and personality Research. R package version 1.7.5. 2017b. Disponível em: https://CRAN.R-project.org/package=psych. Acesso em: 8 set. 2017.

SÃO PAULO (Estado). Secretaria da Educação. *Currículo do Estado de São Paulo*. Ciências da Natureza e suas tecnologias. Ensino Fundamental – Ciclo II e Ensino Médio. São Paulo, 2011a. Disponível em: http://www.educacao.sp.gov.br/a2sitebox/arquivos/documentos/235.pdf. Acesso em: 30 nov. 2018.

SÃO PAULO (Estado). Secretaria da Educação. *Currículo do Estado de São Paulo*. Ciências Humanas e suas tecnologias. Ensino Fundamental – Ciclo II e Ensino Médio. São Paulo, 2011b. Disponível em: http://www.educacao.sp.gov.br/a2sitebox/arquivos/documentos/236.pdf. Acesso em: 30 nov. 2018.

SÃO PAULO (Estado). Secretaria da Educação. *Currículo do Estado de São Paulo*. Linguagens, códigos e suas tecnologias. Ensino Fundamental – Ciclo II e Ensino Médio. São Paulo, 2011c. Disponível em: http://www.educacao.sp.gov.br/a2sitebox/arquivos/documentos/237.pdf. Acesso em: 30 nov. 2018.

SÃO PAULO (Estado). Secretaria da Educação. *Currículo do Estado de São Paulo*. Matemática e suas tecnologias. São Paulo, 2011d. Disponível em: http://www.educacao.sp.gov.br/a2sitebox/arquivos/documentos/238.pdf. Acesso em: 30 nov. 2018.

SCHILLING, S. G. The role of psychometric modeling in test validation: an application of multidimensional item response theory. *Measurement: Interdisciplinary Research & Perspective*, Philadelphia, USA, v. 5, n. 2, p. 93-106, Aug. 2009.

THIMOTY, A. B. *Confirmatory factor analysis for applied research*. 2. ed. New York: The Guilford Press, 2015.

VAN DER LINDEN, W. J. *Handbook of item response theory*. Volume one: models. New York: CRC Press, 2016.

VIEIRA, N. A. *As provas das quatro áreas do Enem vista como prova única na óptica de modelos da teoria da resposta ao item uni e multidimensionais*. 2016. Dissertação (Mestrado em Métodos e Gestão em Avaliação) – Universidade Federal de Santa Catarina, Florianópolis, 2016. Disponível em: https://pergamum.ufsc.br/pergamum/biblioteca/index.php. Acesso em: 23 jul. 2017.

WEAKLIEM, D. L. A critique of the bayesian information criterion for model selection. *Sociological Methods & Research*, London, v. 27, n. 3, p. 359-97, Feb. 1999.

NOTA: As pesquisadoras Ligia Maria Vettorato Trevisan e Tânia Cristina Arantes de Macedo Azevedo foram as autoras do tema e proposta do artigo. Em especial, a Profa. Ligia acompanhou pontualmente todo o trabalho e, juntamente com o Prof. Guaracy Tadeu Rocha, desenvolveu a interpretação pedagógica da escala de proficiência. Os professores Pedro Alberto Barbetta e Dalton Francisco de Andrade realizaram as análises estatísticas que compõem a metodologia de desenvolvimento do presente artigo.

Recebido em: 2 MARÇO 2018
Aprovado para publicação em: 2 JANEIRO 2019