European Journal of Education Studies
ISSN: 2501 - 1111
ISSN-L: 2501 - 1111
Available on-line at: www.oapub.org/edu

doi: 10.5281/zenodo.2527904 Volume 5 | Issue 8 | 2018

ANÁLISE DA FIDEDIGNIDADE COMPOSTA DOS ESCORES DO ENEM POR MEIO DA ANÁLISE FATORIAL DE ITENS[i]

Cristiano Mauro Assis Gomes[1],
Hudson Fernandes Golino[2],
Alexandre José de Souza Peres[3ii]
[1]Federal University of Minas Gerais,
Universidade Federal de Minas Gerais, Departamento de Psicologia,
gabinete 4036, Campus Pampulha, Av. Antônio Carlos, 6627,
Pampulha, Belo Horizonte, Minas Gerais, Brasil
[2]University of Virginia, USA
[3]Federal University of Mato Grosso do Sul, Brazil

Resumo:
O Exame Nacional do Ensino Médio (ENEM) é composto por quatro domínios: Ciências da Natureza, Ciências Humanas, Linguagens e Matemática. O presente estudo busca avaliar a fidedignidade dos escores de cada domínio por meio das seguintes estratégias combinadas: emprego da análise fatorial de itens em duas vertentes, sendo elas a análise fatorial confirmatória e a modelagem exploratória por equação estrutural; o uso da modelagem bifatorial; e o cálculo da fidedignidade composta para os quatro escores do ENEM e o escore geral de desempenho escolar, obtidos pelas estratégias anteriores. Os dados analisados são referentes aos escores dos estudantes que fizeram a prova de 2011. Os resultados apontaram uma fidedignidade satisfatória para o escore geral e os escores nos domínios de Matemática, Ciências da Natureza e Linguagens, enquanto o domínio de Ciências Humanas apresentou uma fidedignidade insatisfatória. Implicações são discutidas.

Palavras-chave: Exame Nacional do Ensino Médio (ENEM); análise fatorial confirmatória; modelagem exploratória por equações estruturais; modelo bifactor; fidedignidade composta.

Abstract:
The National Exam of Upper Secondary Education (Exame Nacional do Ensino Médio [ENEM]) consists of four domains: Natural Sciences, Humanities, Languages, and Mathematics. The present study aims to evaluate the reliability of the scores of each domain through the following strategies: use of item factor analysis via confirmatory

[i] ANALYSIS OF COMPOSITE TRUST SCORES OF ENEM BY FACILITY ANALYSIS OF ITEMS
[i] Correspondence: email alexandre.peres@gmail.com

factor analysis and exploratory structural equation modeling; bifactor model; and the calculation of the composite reliability for the scores from the four domains, plus the general performance factor, all of them obtained by the previous analyzes. The data of the study are composed of the scores of the students who took the 2011 ENEM. The results showed a satisfactory reliability for the scores of the general factor, and for the domains of Mathematics, Nature Sciences, and Languages, while the Humanities domain showed an unsatisfactory reliability. Implications are discussed.

Keywords: National Exam of Secondary Education (ENEM); confirmatory factor analysis; exploratory structural equation modeling; bifactor model; composite reliability.

1. Introdução

O Exame Nacional do Ensino Médio (ENEM) é um teste educacional padronizado organizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), uma autarquia federal vinculada ao Ministério da Educação brasileiro. Desde 2009, o ENEM é administrado anualmente a milhões de estudantes brasileiros que concluíram a Educação Básica, uma vez que é utilizado como critério na seleção para o ingresso na Educação Superior (Inep, 2012a, 2012b, 2016, 2018).

O ENEM é composto por quatro sub-testes, com 45 itens cada um, que resultam em quatro escores (Inep, 2012b): Linguagens, códigos e suas tecnologias (LC); Matemática e suas tecnologias (MT); Ciências da Natureza e suas tecnologias (CN); e Ciências Humanas e suas tecnologias (CH). Dada a importância do ENEM para a vida prática de milhões de pessoas, é necessário que o teste seja investigado quanto à diferentes aspectos relacionados a validade e a fidedignidade dos escores que produz (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing, 2014). Na literatura, é possível encontrar estudos preocupados com a questão da validade de conteúdo do ENEM (e.g., Almeida & Sanchez, 2016; Luna & Marcuschi, 2015; Silveira, Barbosa, & Silva, 2015) ou quanto aos usos de seus resultados como indicador de qualidade educacional (e.g., Andrade & Soida, 2015; Haguette, Pessoa, & Vidal, 2016; Travitzki, 2013). No entanto, ainda são raros os estudos sobre as propriedades psicométricas do Exame, como a fidedignidade. O próprio Inep não divulga em seus relatórios os parâmetros psicométricos dos itens e dos subtestes quem compõem o ENEM, incluindo a fidedignidade dos escores.

Usualmente a estimativa da fidedignidade dos escores de testes é realizada por meio do cáculo do alfa de Cronbach. É gigantesca a literatura que emprega seu cálculo e baseia-se nesse índice para aferir a fidedignidade, inclusive na área da educação (Taber, 2017; Travitzki, 2017). No entanto, recentemente esse coeficiente vem sendo fortemente criticado como indicador da fidedignidade de escores provenientes de modelos multidimensionais (Schmitt, 1996; Sijtsma, 2009; Revelle & Zinbarg, 2009). O principal motivo dessa crítica é o pressuposto do alfa de Cronbach relativo à tau-equivalência: as

variáveis observáveis devem apresentar a mesma carga fatorial junto às variáveis latentes estimadas (Graham, 2006). Esse pressuposto é irreal na maior parte das condições empíricas, pois dificilmente as variáveis observáveis apresentam a mesma carga fatorial nos diferentes fatores de uma estrutural multi-fatorial. Por exemplo, em um teste que possui uma estrutura de três fatores, essa pressuposição implica que um determinado item obrigatoriamente possua a mesma carga fatorial nos três fatores. Ao contrário desse pressuposto, o que normalmente ocorreria é um item apresentar diferentes cargas fatoriais nos três fatores.

Por postular a tau-equivalência, o alfa de Cronbach não insere as cargas fatoriais dentro das estimativas do cálculo da fidedignidade dos escores de um teste. No entanto, uma série de índices vêm sendo propostos como alternativas ao alfa de Cronbach, visando superar essa limitação para uma estimativa mais correta da fidedignidade de testes multidimensionais, como é o caso do ENEM, que possui quatro escores referentes aos domínios. A fidedignidade composta (i.e., *composite reliability*) é um desses índices (Revelle & Condon, 2018; Valentini et al., 2015).

$$\text{confiabilidade composta} = \frac{(\sum \text{betas})^2}{(\sum \text{betas})^2 + \sum \text{erros}}$$

Nela, os betas relacionados à uma determinada variável latente e às variáveis observáveis são somados e essa soma é elevada ao quadrado: $(\sum \text{betas})^2$. Os erros de cada variável observável também são somados e representam a variância não explicada: $\sum \text{erros}$. Por exemplo, se uma variável latente possui um beta de 0,5 em relação a uma determinada variável observável, o erro será de 1 menos 0,5 elevado ao quadrado, indicando o valor de 0,75, ou seja, 75% da variância desta variável observável não é explicada pela variável latente. Esse valor de 0,75 é somado ao outros valores de erro das outras variáveis observáveis em relação a esta variável latente. Os betas somados e elevados ao quadrado são somados aos erros $(\sum \text{betas})^2 + \sum \text{erros}$. Este resultado é usado para servir de denominador, o que define a fidedignidade composta (Fornell & Larcker, 1981). Hair, Black, Babin e Anderson (2013) sugerem um ponto de corte de 0,70 para a fidedignidade composta, valor esse semelhante ao proposto para o alfa de Cronbach.

Conjuntamente ao problema do postulado da tau-equivalência, a literatura psicométrica tem discutido recentemente o uso do modelo bifatorial (Cucina & Bile, 2017; Reise, Bonifay, & Haviland, 2018; Reise, Moore, & Haviland, 2010), devido à necessidade de se separar o efeito de variáveis latentes de ordem mais alta sobre variáveis latentes de mais baixa ordem. Explicando por meio do exemplo do ENEM, o escore do domínio de MT, CH, CN ou LC pode ser influenciado por um fator geral de desempenho, de modo que parte dos escores nos domínios seja explicada, de fato, por esse escore geral. Qual a importância disso para a fidedignidade dos escores? Se um educador pretende obter informação do escore de uma determinada escola no domínio de MT, ele não quer que esse escore esteja enviesado pelo escore geral. Ou seja, ele deseja que o escore do domínio de MT não tenha a influência do escore geral, para que

ele possa ter uma estimativa mais precisa do escore de MT especificamente. Em outros termos, se alguém quer medir o peso do seu corpo e segura uma barra que pesa 10 quilos, será necessário soltar essa barra para que o peso medido não seja uma mistura de seu peso corporal com o peso da barra.

O que a literatura em psicometria tem mostrado atualmente é que os fatores de mais alta ordem, como é o caso de um escore geral, têm influência sobre os escores dos fatores mais específicos e vice-versa. No entanto, é possível ponderar essa influência, da mesma maneira que a pessoa pode soltar a barra e mensurar de forma mais confiável seu peso corporal. A solução? Modelos bifatoriais, porque eles têm a capacidade de separar o efeito dos diferentes fatores junto aos itens de um teste, propiciando escores que não apresentam a influência das outras variáveis latentes do modelo.

O presente estudo pretende avaliar a fidedignidade do ENEM por meio da utilização dos escores de acerto e erro (1 e 0) dos 180 itens do teste, aplicando um modelamento bifatorial na matriz de correlação desses itens, e calculando a fidedignidade composta dos domínios, a partir dos betas obtidos por essa análise. A diferença do estudo atual em relação ao estudo anterior envolve o número bem maior de variáveis observáveis. Essa é uma vantagem considerável, pois a fidedignidade é bastante influenciada pelo número de variáveis observáveis presentes em um determinado fator. Quanto mais itens carregam em um fator, maior tende a ser a estimativa da fidedignidade de seu escore.

2. Método

2.1 Participantes

Analisou-se os escores de 66.880 estudantes que participaram do ENEM de 2011 e completaram especificamente os cadernos 120, 124, 125 e 129 (Inep, 2012a). As características demográficas dos participantes eram as seguintes: média de idade de 21,48 anos (*DP* = 7,12); 53,3% do sexo feminino; 50,5% autodeclarados brancos, 10,4% pretos, 33,4% pardos, 2,5% amarelos, 0,5% indígenas, e 2,8 não declararam a cor da pele ou etnia.

2.2 Instrumento

A prova de 2011do ENEM foi composta por 180 itens (Inep, 2012b), sendo que cada subteste (i.e., domínio) apresentava 45 itens: MT (Matemática), LC (Linguagens e seus códigos), CN (Ciências da Natureza) e CH (Ciências Humanas).

2.3 Procedimentos

Os microdados do ENEM de 2011 estão disponíveis publicamente para *download* (Inep, 2012a). O *download*, a extração, a importação e o tratamento inicial dos dados foi realizada por meio do pacote *ENEM* (Golino, 2014), desenvolvido para o software *R* (R Core Team, 2013). O tratamento inicial consistiu na exclusão dos participantes ausentes nas provas e a correção dos vetores de resposta de acordo com o gabarito.

2.4 Análise de Dados

A análise dos dados foi realizada por meio de uma integração entre os softwares R (R Core Team, 2013) e Mplus (Muthén & Muthén, 2014). Os modelos foram construídos no R, por meio do pacote *MplusAutomation* (Hallquist & Wiley, 2014), enquanto os cálculos e estimativas foram realizadas no Mplus. O ajuste dos modelos aos dados foi verificado por meio dos índices RMSEA (*root mean square error of approximation*), CFI (*the comparative fit index*) e TLI (Tucker-Lewis Index). Um bom ajuste aos dados é indicado se o RMSEA é igual ou inferior a 0,05, um CFI e TLI igual ou superior a 0,95 (Bentler, 1990; Browne & Cudeck, 1993; Hu & Bentler, 1999; Schumacker & Lomax, 2004).

Não é improvável que diferentes domínios tenham alguma influência em um mesmo item do ENEM, a despeito desse item ter sido elaborado inicialmente para quantificar o escore de um domínio específico. Essa condição é possível em função das características dos itens, que são elaborados a partir de um princípio de interdisciplinaridade. Por exemplo, alguns itens de matemática podem também requerer a ativação do domínio de ciências da natureza, e vice-versa. Assim, buscando maximizar a estimativa da fidedignidade dos escores dos quatro domínios do ENEM, foi realizado um modelamento por equação estrutural exploratório (ESEM – *Exploratory Structural Equation Modeling*) e que permite uma análise exploratória das cargas fatoriais dos domínios em relação a todos os itens da prova (Gomes, Almeida, & Núñez, 2017). Esse modelo envolve uma modelagem híbrida, onde os fatores são determinados de forma confirmatória, a priori, e as cargas fatoriais dos fatores nos itens são determinados a posteriori, de forma exploratória. O modelamento por ESEM envolveu a determinação bifatorial de cinco variáveis latentes: um fator geral de desempenho escolar e quatro fatores referentes aos domínios do ENEM. Todas as variáveis latentes se relacionavam de forma ortogonal. Chamamos este modelo de Modelo 1.

Apesar de a ESEM ser uma técnica vantajosa para descobrir itens que carregam nos diferentes domínios e que não foram elaborados *a priori* para sua quantificação, essa técnica permite a presença de cargas fatoriais negativas e elas não facilitam a estimativa da fidedignidade composta dos domínios. Em função disso, uma análise fatorial confirmatória (CFA) foi realizada, tendo por base o modelo anterior. No entanto, todas as cargas negativas do modelo anterior foram constrangidas para o valor de zero neste segundo modelo. Nesse caso, manteve-se as relações entre os domínios e os itens que apresentaram uma carga fatorial igual ou superior a 0,10. Chamamos este modelo de Modelo 2.

3. Resultados

O ajuste do Modelo 1, estimado por meio de ESEM, mostrou-se muito bom (χ^2[15220] = 72394,860; CFI = 0,988; TLI = 0,987; RMSEA = 0,007). A Tabela 1 apresenta as cargas fatoriais dos quatro domínios (i.e., LC, MT, CN e CH) e do fator de desempenho escolar geral (DEG). A maioria dos itens apresenta a maior carga fatorial no domínio de desempenho escolar geral e, a despeito de haver cargas fatoriais negativas nos diferentes atores, elas normalmente apresentam carga inferior ao valor absoluto de 0,10.

O domínio de CN apresentou o maior número de cargas fatoriais acima de 0,10 em outros itens que não naqueles elaborados para sua própria quantificação.

Tabela 1: Cargas fatoriais dos itens dos domínios teóricos em relação aos cinco fatores do Modelo 1

Domínios teóricos	Fatores do Modelo 1	Cargas fatoriais					Total de itens
		Maior	Menor	Média*	Desvio-Padrão*	Mediana*	
LC	DEG	0,72	0,13	0,45	0,14	0,45	45
	CN	0,20	-0,26	0,09	0,06	0,09	45
	MT	0,04	-0,12	0,04	0,03	0,04	45
	LC	0,40	-0,08	0,14	0,09	0,12	45
	CH	0,08	-0,41	0,07	0,09	0,04	45
MT	DEG	0,60	0,00	0,40	0,15	0,44	45
	CN	0,48	-0,11	0,18	0,11	0,21	45
	MT	0,41	-0,13	0,18	0,12	0,17	45
	LC	0,06	-0,12	0,04	0,03	0,03	45
	CH	0,12	-0,09	0,03	0,02	0,03	45
CN	DEG	0,63	-0,06	0,34	0,17	0,32	45
	CN	0,49	-0,24	0,14	0,11	0,11	45
	MT	0,16	-0,09	0,05	0,04	0,05	45
	LC	0,12	-0,16	0,05	0,04	0,04	45
	CH	0,19	-0,08	0,05	0,05	0,04	45
CH	DEG	0,73	0,06	0,43	0,15	0,45	45
	CN	0,30	-0,26	0,09	0,07	0,07	45
	MT	0,07	-0,15	0,07	0,04	0,07	45
	LC	0,13	-0,23	0,05	0,05	0,04	45
	CH	0,21	-0,07	0,06	0,05	0,05	45

Legenda: LC (Linguagens e Códigos); MT (Matemática); CN (Ciências da Natureza); CH (Ciências Humanas); DEG (Desempenho Escolar Geral).

* Estatísticas calculadas considerando os valores absolutos das cargas fatoriais.

A Figura 1 mostra o padrão de distribuição das cargas fatoriais dos itens por fator do modelo ESEM bifatorial. No quadrante superior esquerdo encontram-se todas as cargas fatoriais com valores iguais ou superiores à 0,10. Já no quadrante superior direito, encontram-se os itens com as cargas fatoriais iguais ou superiores à 0,30, e no quadrante inferior esquerdo os itens com cargas iguais ou superiores à 0,50. Por último, no quadrante inferior direito encontram-se os itens com cargas iguais ou superiores à 0,70.

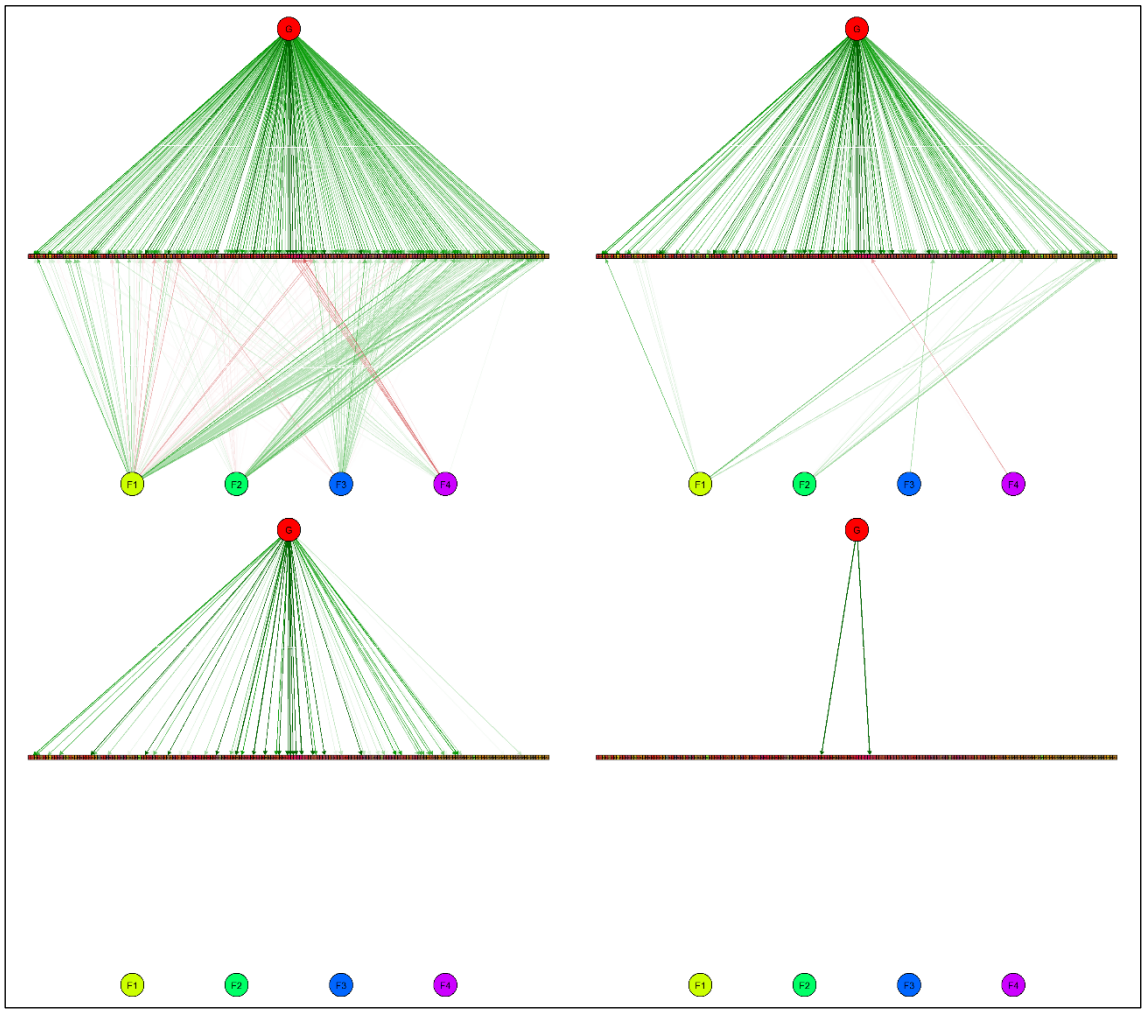

Figura 1: Padrão de distribuição das cargas fatoriais padronizadas dos itens por fator no Modelo 1

O Modelo 2, estimado por meio de CFA em que todas as cargas negativas do modelo anterior foram constrangidas para o valor de zero, apresentou adequado grau de ajuste aos dados (χ^2[15795] = 179708,023; CFI = 0,965; TLI = 0,964; RMSEA = 0,012). A Tabela 2 mostra os betas dos diferentes domínios em relação aos itens da prova.

Tabela 2: Cargas fatoriais dos itens dos domínios teóricos em relação aos cinco fatores do Modelo 2

Domínios Teóricos	Fatores do Modelo 2	Cargas fatoriais					Total de itens
		Maior	Menor	Média	Desvio-Padrão	Mediana	
LC	DEG	0.74	0.11	0.44	0.14	0.44	45
	CN	0.20	0.15	0.18	0.02	0.18	4
	MT						
	LC	0.45	0.11	0.26	0.08	0.27	28
	CH						
MT	DEG	0.58	0.00	0.37	0.14	0.41	45
	CN	0.52	0.17	0.27	0.08	0.26	31
	MT	0.41	0.12	0.29	0.09	0.30	29
	LC						0

	CH	0.09					1
	DEG	0.63	0.00	0.34	0.18	0.31	45
	CN	0.50	0.09	0.22	0.10	0.20	23
CN	MT	0.15	0.06	0.11	0.05	0.11	2
	LC	0.18					1
	CH						0
	DEG	0.73	0.04	0.43	0.15	0.44	45
	CN	0.32	0.12	0.20	0.06	0.18	8
CH	MT						0
	LC	0.17	0.17	0.17	0.00	0.17	2
	CH	0.39	0.10	0.21	0.08	0.18	9

Legenda: LC (Linguagens e Códigos); MT (Matemática); CN (Ciências da Natureza); CH (Ciências Humanas); DEG (Desempenho Escolar Geral).

A Figura 2 mostra o padrão de distribuição das cargas fatoriais dos itens por fator do modelo bifatorial final, verificado via análise fatorial confirmatória. No quadrante superior esquerdo encontram-se todas as cargas fatoriais com valores iguais ou superiores à 0,10. Já no quadrante superior direito, encontram-se os itens com as cargas fatoriais iguais ou superiores à 0,30, e no quadrante inferior esquerdo os itens com cargas iguais ou superiores à 0,50. Por último, no quadrante inferior direito encontram-se os itens com cargas iguais ou superiores à 0,70.

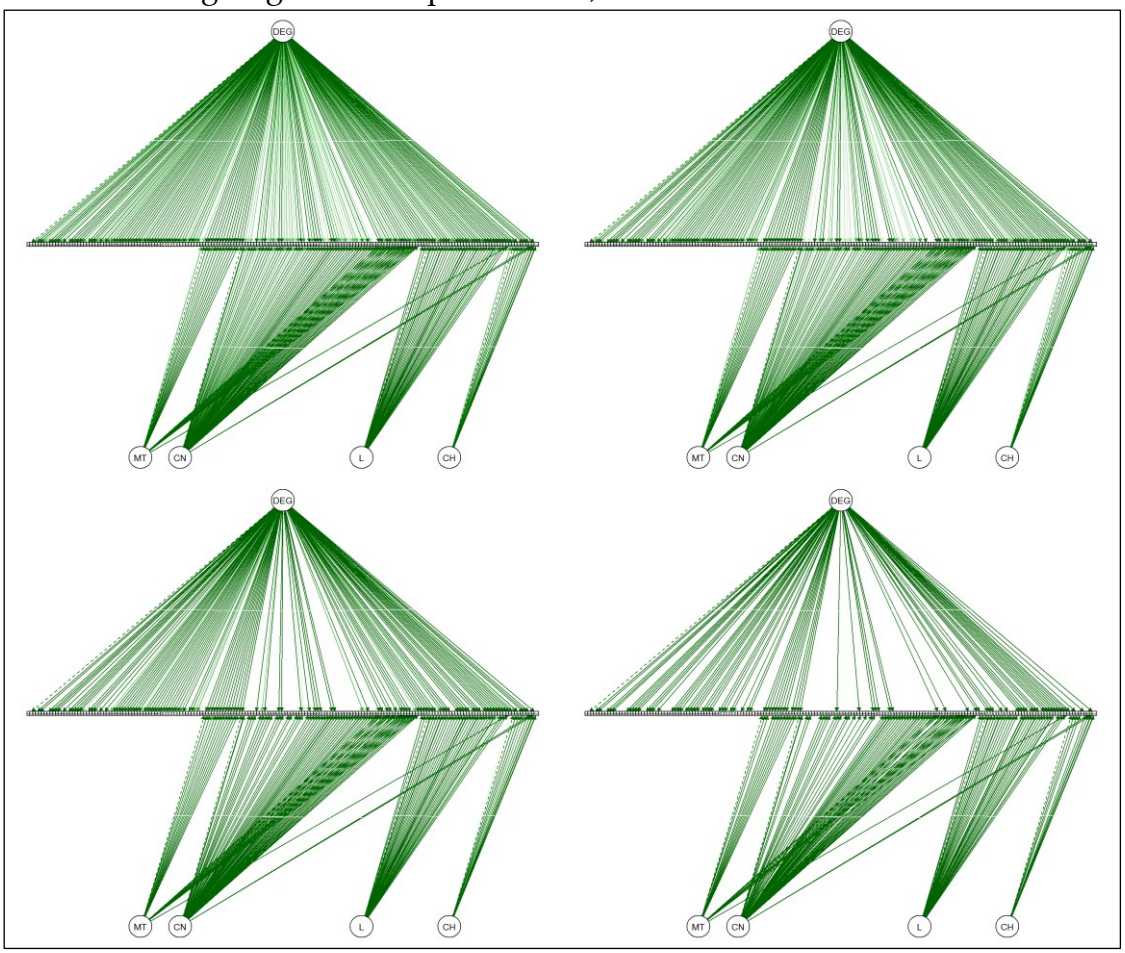

Figura 2: Padrão de distribuição das cargas fatoriais padronizadas dos itens por fator no Modelo 2

Os betas apresentados na Tabela 2 foram utilizados para o cálculo da fidedignidade composta de cada um domínios. Os coeficientes de fidedignidade composta dos fatores do Modelo 2 são apresentados na Tabela 3, em conjunto com os coeficientes alfa desse mesmo modelo e do modelo de fatores não-correlacionados, adotado atualmente pelo Inep no ENEM. Pode-se observar que os coeficientes alfa são diferentes para os fatores nos dois modelos avaliados, sendo maiores no modelo de fatores não-correlacionados. Além disso, a estimativa por meio do coeficiente de fidedignidade composta revela uma precisão menor nos quatro domínios específicos (i.e., CH, CN, LC e MT), mas maior para o fator DEG. Essas diferenças são explicadas justamente pelas estratégias analíticas adotadas neste estudo. Ao inserir as cargas fatoriais nas estimativas da fidedignidade composta, os coeficientes sofreram decréscimo em relação ao alfa, que assume o postulado da tau-equivalência. O fator DEG apresenta maior fidedignidade justamente por também apresentar maior média nas cargas fatoriais de seus itens. As diferenças nas estimativas também são explicadas pela adoção do modelo bifatorial, que separou o efeito do fator geral DEG sobre os quatro domínios específicos. Assim, no caso do alfa, as estimativas diminuíram para os fatores que perderam itens. Já no caso da fidedignidade composta, as estimativas foram menores para aqueles fatores com menor média nas cargas fatoriais dos itens.

Tabela 3: Coeficientes de fidedignidade do modelo de fatores não-correlacionados e do modelo bifatorial final (Modelo 2)

	Modelo de fatores não-correlacionados		Modelo 2		
Fatores	**Alfa**	**Itens**	**Alfa**	**Fidedignidade composta**	**Itens**
CH	0,916	45	0,576	0,298	6
CN	0,872	45	0,915	0,805	66
LC	0,849	45	0,839	0,678	31
MT	0,924	45	0,922	0,720	31
DEG			0,968	0,972	177

Legenda: LC (Linguagens e Códigos); MT (Matemática); CN (Ciências da Natureza); CH (Ciências Humanas); DEG (Desempenho Escolar Geral).

4. Discussão e Conclusão

O ENEM é uma avaliação educacional que verifica o desempenho de estudantes em quatro grandes domínios: Ciências Humanas, Ciências da Natureza, Linguagens e Códigos e Matemática. Todos os anos são divulgados os escores dos estudantes em cada domínio. Um passo importante para se verificar a qualidade do ENEM enquanto avaliação educacional envolve verificar se esses escores são válidos e confiáveis. Como apontado anteriormente, a literatura psicométrica tem discutido recentemente o uso do modelo bifatorial, que permite separar o efeito de variáveis latentes de alto nível sobre variáveis latentes de mais baixo nível (Reise, Moore, & Haviland, 2010). A vantagem de se investigar a validade dos escores do ENEM por meio da aplicação de um modelo bifatorial estimado seja via modelagem de equações estruturais exploratórias ou via análise fatorial confirmatória diz respeito à possibilidade de verificar o impacto de um

fator geral educacional (denominado desempenho escolar geral, ou DEG) nos fatores específicos que representam os quatro domínios do ENEM.

Os resultados do presente estudo apontam que o modelo bifatorial ESEM (i.e., Modelo 1) apresenta um bom ajuste aos dados (χ^2[15220] = 72394,860; CFI = 0,988; TLI = 0,987; RMSEA = 0,007). No entanto, apenas o fator geral escolar apresenta uma quantidade relevante de itens com cargas padronizadas superiores à 0,30 (ver Figura 1). Esse modelo, apesar de apresentar um bom ajuste aos dados, permite a presença de cargas fatoriais negativas que dificultam a estimativa da fidedignidade composta dos domínios. Por esse motivo, foi realizada uma análise fatorial confirmatória tendo por base o Modelo 1, em que foram mantidas apenas as relações entre os domínios e os itens que apresentaram uma carga fatorial igual ou superior a 0,10. Esse modelo (i.e., Modelo 2) também apresentou um bom ajuste aos dados (χ^2[15795] = 179708,023; CFI = 0,965; TLI = 0,964; RMSEA = 0,012), e o número de itens com cargas fatoriais superiores à 0,30 nos domínios aumentou (ver Figura 2).

A fidedignidade composta calculada à partir desse modelo final (i.e., Modelo 2) apontou que o Domínio Escolar Geral (DEG) possui uma fidedignidade muito elevada (0,97), enquanto os domínios de Ciências da Natureza (0,80) e Matemática (0,72) apresentaram uma fidedignidade adequada. No entanto, o domínio de Linguagens e Códigos apresentou uma fidedignidade limítrofe (0,678), enquanto o domínio de Ciências Humanas apresentou uma fidedignidade muito baixa (0,298).

A literatura tem apontado que coeficientes de fidedignidade de escores de variáveis latentes com valores iguais ou acima de 0,70, são indicativos de escores confiáveis (Hogan, 2013). Por sua vez, valores entre 0,60 a 0,69 são valores aceitáveis para fins de pesquisa, podendo ser considerados como uma fidedignidade limítrofe. No entanto, no contexto de avaliações *high stake* como o ENEM, de acordo com os critérios revisados por Hogan (2013), são esperados coeficientes de 0,95. Considerando esses critérios, apenas o fator DEG do modelo bifatorial apresentaria fidedignidade aceitável.

Os resultados deste estudo também têm impacto sobre questões relativas ao modelo teórico subjacente ao ENEM, expresso em sua matriz de referência (Inep, 2012b), e para o uso prático dos escores. Nesse sentido, duas perspectivas devem ser destacadas. Primeiramente, como relatado, os resultados indicam a presença do fator geral DEG. Do ponto de vista desenvolvimental, é possível conjecturar que esse construto é operacionalizado por itens que exigem dos estudantes uma estruturação interdisciplinar das diferentes disciplinas trabalhadas ao longo da Educação Básica. Ou seja, o ENEM de fato parece exigir dos estudantes a mobilização de habilidades, competências e conhecimentos de forma integrada e multidisciplinar para a resolução de problemas.

O segundo aspecto a ser destacado diz respeito ao desafio de operacionalizar os construtos teóricos atualmente presentes no *framework* do ENEM (Inep, 2012b): quatro domínios teóricos (i.e., Linguagens, códigos e suas tecnologias, Matemática e suas tecnologias, Ciências da Natureza e suas tecnologias, e Ciências Humanas e suas tecnologias), compostos por uma série de construtos de ordem inferior, chamados de Competências de Área que, por sua vez, são compostas por habilidades. Neste estudo, a

consistência interna resultou em fidedignidade inaceitável para os domínios teóricos, com exceção do DEG (0,97). Ou seja, a operacionalização da matriz de referência por meio dos itens que compõem as provas do ENEM parece não ser capaz de avaliar especificamente cada domínio isoladamente. No entanto, os quatro domínios contribuem para a formação do DEG. Essas informações são indicativos importantes de que o ENEM consiste em um modelo multidimensional, o que não deixa de ser algo contraditório, pois apenas o fator geral apresenta uma confiabilidade condizente para um teste educacional *high stake* de larga-escala.

Em suma, o presente estudo reforça o argumento de que o acréscimo no número de variáveis observáveis tende a aumentar a fidedignidade das dimensões investigadas, embora isso não implique automaticamente em um bom modelo de medida. Por um lado, quando analisa-se a fidedignidade assumindo o pressuposto da tau-equivalência e a ausência de correlação entre os fatores, os índices de fidedignidade são mais altos – embora não satisfatórios para um teste *high stake*. Por outro lado, quando modelamos as mesmas variáveis considerando a presença de um fator geral e multidimensional, verificamos que o poder explicativo dos fatores teóricos diminui e, consequentemente, sua fidedignidade. No entanto, o poder explicativo do modelo é potencializado, tanto pela presença de um fator geral com alta fidedignidade, quanto pela discriminação da real influência dos fatores teóricos no modelo. Considerando a relevância do fator geral e sua influência na estimativa da confiabilidade dos domínios teóricos do ENEM, sugerimos ao INEP que adote em seu modelo analítico a presença do fator geral, assim como investigue a confiabilidade dos escores dos domínios por meio do modelo bifatorial e da confiabilidade composta. O modelo bifatorial e a confiabilidade composta, combinados em uma mesma análise, são uma estratégia mais apropriada e bastante efetiva para se investigar a confiabilidade de escores de variáveis latentes, como é o caso do ENEM.

Referências

Almeida, Marco Antonio Bettine de, & Sanchez, Lívia Pizauro. (2016). ENEM: ferramenta de implementação da Lei 10.639/2003 - Competências para a transformação social? Educação em Revista, 32(1), 79-103. doi: 10.1590/0102-4698141429

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). Standards for educational and psychological testing.

Andrade, Eduardo, & Soida, Ivan. (2015). A qualidade do ranking das escolas de ensino médio baseado no ENEM é questionável. Estudos Econômicos (São Paulo), 45(2), 253-286. doi: 10.1590/0101-4161201545221eai

Bentler, P.M. (1990). Comparative fit indexes in structural models. Psychological Bulletin, 107(2), 238-246. doi: 10.1037/0033-2909.107.2.238

Browne, M.W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In: Bollen, K.A., Long, J.S. (Ed.) Testing structural equation models (pp.136 -162). Newbury Park: Sage.

Cucina, J., & Byle, K. (2017). The bifactor model fits better than the higher-order model in more than 90% of comparisons for mental abilities test batteries. Journal of Intelligence, 5(27), 1-21. doi: 10.3390/jintelligence5030027

Fornell, C., & Larcker, D.F. (1981). Evaluating structural equations models with unobservable variables and measurement error. Journal of Marketing,18(1), 39-50. doi: 10.2307/3151312

Golino, H.F. (2014). ENEM: an implementation of functions to help automatic downloading, importing, cleaning and scoring of the Brazilian's National High School Exam (ENEM). Software não-publicado.

Gomes, C.M.A., Almeida, L.S., & Núñez, J.C. (2017). Rationale and Applicability of Exploratory Structural Equation Modeling (ESEM) in psychoeducational contexts. Psicothema, 29(3), 396-401. doi: 10.7334/psicothema2016.369.

Graham, J.M., 2006. Congeneric and (essentially) tau-equivalent estimates of score reliability. What they are and how to use them. Educational and Psychological Measurement, 66(6), 930-944. doi: 10.1177/0013164406288165

Haguette, A., Pessoa, M.K.M., & Vidal, E.M. (2016). Dez escolas, dois padrões de qualidade. Uma pesquisa em dez escolas públicas de Ensino Médio do Estado do Ceará. Ensaio: Avaliação e Políticas Públicas em Educação, 24(92), 609-636. doi: 10.1590/S0104-40362016000300005

Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2013). Multivariate data analysis. Pearson new international edition (7th ed.). Essex, UK: Pearson Education Limited.

Hallquist, M. & Wiley, J. (2014). MplusAutomation: Automating Mplus Model Estimation and Interpretation. R package. (Version 0.6-3) [Software]. Available from http://CRAN.R-project.org/package=MplusAutomation

Hogan, T.P. (2013). Psychological testing: A practical introduction (3rd ed.). Hoboken, USA: John Wiley & Sons.

Hu, L.T. & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modeling, 6(1), 1-55. doi: 10.1080/10705519909540118

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - Inep. (2016). Overview of the Brazilian education system. Brasília: Inep.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – Inep. (2012a). Microdados do ENEM – 2011. Exame Nacional do Ensino Médio: Manual do Usuário. Retrieved from http://portal.inep.gov.br/web/guest/microdados. Access 17 September 2018.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - Inep. (2012b). Matriz de Referência do ENEM. Retrieved from http://download.inep.gov.br/educacao_basica/enem/downloads/2012/matriz_referencia_enem.pdf. Access 17 September 2018.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - Inep. (2018). ENEM. Retrieved from portal.inep.gov.br/web/guest/enem Access 17 September 2018.

Luna, T.S., & M.,B. (2015). Letramentos literários: o que se Avalia No Exame Nacional Do Ensino Médio? Educação em Revista, 31(3), 195-224. doi: 10.1590/0102-4698135569

Muthén, L.K. & Muthén, B.O. (2014). Mplus users's guide. Los Angeles: Muthén & Muthén.

R. Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Reise, S.P., Bonifay, W., & Haviland, M.G., 2018). Bifactor modelling and the evaluation of scale scores. In P. Irwing, T. Booth, & D. J. Hughes (Eds.). The Wiley Handbook of Psychometric Testing: A multidisciplinary reference on survey, scale and test development (Vols. 1-2) (pp. 677-706). Hoboken, NJ, EUA: John Wiley & Sons.

Reise, S.P., Moore, T.M., & Haviland, M.G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. Journal of Personality Assessment, 92(6), 544–559. doi: 10.1080/00223891.2010.496477

Revelle, W., & Condon, D.M. (2018). Reliability. In P. Irwing, T. Booth, & D. J. Hughes (Eds.). The Wiley Handbook of Psychometric Testing: A multidisciplinary reference on survey, scale and test development (Vols. 1-2) (pp. 709-749). Hoboken, NJ, EUA: John Wiley & Sons.

Revelle, W., & Zinbarg, R.E. (2009). Coefficients Alpha, Beta, Omega, and the GLB: Comments on Sijtsma, doi: 10.1007/S11336-008-9102-Z

Schmitt, N. (1996). Uses and abuses of coefficient Alpha. Psychological Assessment, 8(4), 350-353. doi: 1040-3590/96/$300

Schumacker, R.E., & Lomax, R.G. (2004). A beginner's guide to structural equation modeling. London: Lawrence Erlbaum Associates.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's Alpha. Psychometrika. 74(1):107-120. doi: 10.1007/s11336-008-9101-0

Silveira, F.L., Barbosa, M.C.B., & Silva, R. (2015). Exame Nacional do Ensino Médio (ENEM): Uma análise crítica. Revista Brasileira de Ensino de Física, 37(1), 1101. doi: 10.1590/S1806-11173710001

Taber, K.S. (2017). The use of Cronbach's Alpha when developing and reporting research instruments in science education. Research in Science Education, 1-24. doi: 10.1007/s11165-016-9602-2.

Travitzki, R. (2017). Avaliação da qualidade do Enem 2009 e 2011 com técnicas psicométricas. Estudos em Avaliação Educacional, 28(67), 256-288. doi: 10.18222/eae.v28i67.3910

Valentini, F., Gomes, C.M.A., Muniz, M., Mecca, T.P., Laros, J.A., & Andrade, J.M. (2015). Confiabilidade dos índices fatoriais da Wais-III adaptada para a

população brasileira. Psicologia: Teoria e Prática, 17(2), 123-139. doi: 10.15348/1980-6906/psicologia.v17n2p123-139