OPEN ACCESS

ESCOLA DE CIÊNCIAS DA SAÚDE E DA VIDA

PSICO

Psico, Porto Alegre, v. 51, n. 2, p. 1-10, abr.-jun. 2020
e-ISSN: 1980-8623 | ISSN-L: 0103-5371

http://dx.doi.org/10.15448/1980-8623.2020.2.31145

SEÇÃO: ARTIGO

Fidedignidade dos escores do Exame Nacional do Ensino Médio (Enem)

Reliability of scores from the National Exam of Upper Secondary Education (Enem)

Fiabilidad de los escores del Examen Nacional De La Secundaria Superior (Enem)

Cristiano Mauro Assis Gomes[1]
orcid.org/0000-0003-3939-5807
cristianomaurogomes@gmail.com

Hudson Fernandes Golino[2]
orcid.org/0000-0002-1601-1447
hfg9s@virginia.edu

Alexandre José de Souza Peres[3]
orcid.org/0000-0002-3472-6120
alexandre.peres@gmail.com

Recebido em: 21 jun. 2018.
Aprovado em: 17 jan. 2020.
Publicado em: 4 set. 2020.

Resumo: O Exame Nacional do Ensino Médio (Enem) gera uma pontuação para cada domínio que avalia: matemática, linguagens, ciências da natureza e ciências humanas. Reconhecendo a relevância do exame no acesso ao ensino superior e em outros aspectos da vida prática do estudante brasileiro, o presente estudo investiga a fidedignidade dos escores do Enem nos seus quatro domínios. Utilizou-se como amostra os escores dos estudantes que participaram da edição de 2011 do exame. As análises envolveram a estimação dos parâmetros de um modelo de quatro fatores correlacionados e de um modelo bifatorial por meio de análise fatorial confirmatória, além da estimação da fidedignidade composta e da fidedignidade *omega* dos quatro domínios e do fator geral de desempenho, no caso do modelo bifatorial. Utilizou-se como variáveis observáveis as 30 competências de cada domínio. Os resultados indicaram alta fidedignidade apenas para os escores provenientes do fator geral.

Palavras-chave: exame nacional do ensino médio (Enem), fidedignidade, fidedignidade composta, fidedignidade *omega*; modelo bifactor.

Abstract: The National Exam of Upper Secondary Education (ENEM) generates a score for each domains it assess: mathematics, languages, natural sciences, and humanities. Considering the relevance of the Exam in the access to higher education and in other practical aspects of Brazilian students' life, the present study investigates the reliability of the scores from the four domains. We used as a sample the scores of the students who participated in the 2011 edition of the Exam. The analyzes involved the estimation of the parameters of an oblique four-factor model and a bifactor model using confirmatory factor analysis, as well as the estimation of composite and omega reliability of the four domains and the general performance factor, in the case of the bifactor model. We used the 30 competences of each domain as observable variables. The results indicated high reliability only for the scores from the general factor.

Keywords: national exam of upper secondary education (Enem), reliability, composite reliability, omega reliability, bifactor model.

Resumen: El Examen Nacional de la Secundaria Superior (ENEM) genera un puntaje para cada dominio que evalúa: matemática, lenguage, ciencias naturales y humanidades. Puesta la relevancia del Examen en el acceso a la educación superior y en otros aspectos prácticos de la vida de los estudiantes brasileños, el presente estudio investiga la fiabilidad de los puntajes de los cuatro dominios. Usamos como muestra los puntajes de los estudiantes que participaron en la edición 2011 del Examen. Los análisis incluyeron la estimación de los parámetros de un modelo oblicuo de cuatro factores y un modelo bifactorial utilizando análisis factorial confirmatorio, así como la estimación de la confiabilidad compuesta y omega de los cuatro dominios y el factor de desempeño general, en el caso del modelo bifactor. Usamos las 30 competencias de cada dominio como variables observables. Los resultados indicaron una alta fiabilidad solo para los puntajes del factor general.

Palabras clave: examen nacional de la secundaria superior (Enem), fiabilidad, fiabilidad compuesta, fiabilidad omega, modelo bifactor.

[1] Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, MG, Brasil.
[2] University of Virginia (UVA), Charlottesville, VA, EUA.
[3] Universidade Federal de Mato Grosso do Sul (UFMS), Paranaíba, MS, Brasil.

O Exame Nacional do Ensino Médio (Enem) é um teste educacional padronizado brasileiro criado em 1998 pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), autarquia federal vinculada ao Ministério da Educação. Após passar por uma grande reformulação em 2009, o exame é atualmente utilizado, especialmente, para fins de seleção de candidatos para o acesso ao ensino superior, entre outros objetivos (Inep, 2012a, 2018). O modelo atual do Enem produz uma nota para cada um dos domínios que o teste busca avaliar: 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). Esses escores possuem elevada importância social, pois tanto têm influência na vida acadêmica e profissional de muitos estudantes que finalizam o ensino médio, na medida em que diversas universidades públicas e privadas utilizam esses escores para seu processo seletivo, quanto trazem informações para a formulação de diagnósticos educacionais sobre o nível dos estudantes brasileiros no ensino médio em diferentes domínios escolares, podendo influenciar, por sua vez, políticas públicas na área da educação.

Avaliar a qualidade de instrumentos educacionais, inspecionando sua validade e confiabilidade é um procedimento capital (Gomes & Borges, 2008; Gomes, 2010; Pires & Gomes, 2017, 2018), principalmente no caso de provas educacionais de larga-escala de alto impacto e relevância social. Devido à sua importância, podemos encontrar uma quantidade razoável de estudos sobre a validade do Enem, tanto em termos de sua validade externa o (e.g., Gomes & Borges, 2009; Gomes & Jelihovschi, 2019) quanto de sua validade interna (e.g., Gomes, Golino, & Peres, 2016). No entanto, encontramos uma lacuna de estudos que avaliem a confiabilidade dos escores do Enem, considerando-se o estado da arte dos índices de confiabilidade, como é o caso da fidedignidade composta e do *omega* (Gomes, Golino, & Peres, 2018; McDonald, 1999; Raykov, 2001; Revelle & Condon, 2018; Valentini et al., 2015).

Os escores gerados pelo Enem necessitam ser confiáveis e a confiabilidade (i.e., fidedignidade ou precisão) em questão refere-se a um nível máximo aceitável da presença de erro nesses escores. Escores com elevado teor de erro geram informações com excessivo ruído e, portanto, não são dignos de confiança. O grau de fidedignidade ou a quantidade de erro permitido em um escore depende de algumas convenções arbitrárias. Por exemplo, é aceitável imaginar que uma balança de peso de um adulto de 90 quilos possua um erro de 200 gramas, sem comprometer a qualidade do escore obtido. Já uma balança com um erro de 200 gramas pode ser comprometedora quando se pesa um bebê prematuro com peso de um quilograma. Nesse caso, esse erro é insustentável e o escore obtido não é confiável.

A literatura psicométrica tem sugerido que indicadores de fidedignidade de escores de variáveis latentes com valores iguais ou acima de 0,70, em uma escala que vai de zero a um, 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 contexto de avaliações *high stake* (i.e., testes cujos escores são utilizados para decisões importantes), como o Enem, de acordo com os critérios revisados por Hogan (2013), são esperados coeficientes de 0,95. Esses valores são válidos tanto para a estimativa da consistência interna dos itens de uma variável latente, mensurada pelo alfa de Cronbach, quanto para a estimativa da consistência interna dos itens pela fidedignidade composta (*composite reliability*), e para a estimativa da fidedignidade da separação de pessoas e da fidedignidade da separação dos itens, esses dois últimos estimados pelos modelos Rasch (Golino, Gomes, Amarantes, & Coelho, 2015; Revelle & Condon, 2018; Valentini et al., 2015).

A fidedignidade de um teste educacional depende de um conjunto de fatores como, por exemplo, a capacidade de seus itens em aferir as habilidades de estudantes com baixo, médio e alto desempenho e o grau em que seus itens estão relacionados às variáveis latentes alvo (i.e., que deveriam mensurar de acordo com o

referencial teórico), entre outros. Nesse contexto, uma preocupação atual da literatura psicométrica está relacionada a como estimar corretamente a fidedignidade de instrumentos que possuem múltiplos escores, principalmente quando possuem um grau de correlação entre si. Essa preocupação tem sua origem na estimação da fidedignidade dos escores de variáveis latentes específicas, por meio do controle do efeito de uma variável latente geral, assim como na estimação da fidedignidade de uma variável latente geral por meio do controle dos efeitos de variáveis latentes específicas (Reise, 2012; Rios & Wells, 2014).

Um exemplo para ilustrar essa questão é encontrado na literatura psicométrica sobre inteligência. Carroll (1993) apresentou evidências sólidas de que o desempenho das pessoas em todo e qualquer teste de inteligência obrigatoriamente deveria ser explicado por pelo menos três variáveis latentes diferentes. Uma dessas variáveis latentes seria uma habilidade cognitiva específica, relacionada contextualmente com as tarefas dos itens do próprio teste. Em função da proximidade dessa variável latente com o tipo de processo cognitivo ou estímulo presente nas tarefas dos itens, Carroll (1993) a chamou de variável de primeiro nível ou extrato. A segunda dessas variáveis latentes, que explicaria um teste qualquer, seria uma habilidade cognitiva mais ampla não necessariamente tão acoplada ao tipo de processo ou domínio do teste, mas que se vincularia a ele de forma mais abrangente. À época, Carroll (1993) identificou a presença de oito habilidades cognitivas com esse caráter, e as chamou de habilidades de segundo nível ou extrato. Por último, a terceira variável latente a explicar um teste qualquer é a variável latente que ele chamou de fator geral de inteligência. Ele deu esse nome à única variável latente identificada por ele em um nível ainda mais amplo que as variáveis latentes de segundo nível. Por ser a única identificada no terceiro nível, ela teve o nome de fator geral, presente na explicação do desempenho das pessoas em todos os testes de inteligência.

Os argumentos de Carroll (1993) foram contundentes e ele chamou atenção para o fato de que a fidedignidade dos testes de inteligência depende da análise dos efeitos dessas diferentes variáveis latentes sobre o desempenho das pessoas nos mesmos. Por exemplo, o desempenho das pessoas em um teste de raciocínio indutivo é explicado, segundo o modelo de Carroll (2003), tanto pela variável latente de primeiro nível chamada raciocínio indutivo, quanto pela variável latente de segundo nível chamada de inteligência fluida, assim como pela variável latente de terceiro nível, o fator geral de inteligência. Nesse modelo, para estimar de forma correta a fidedignidade do escore de raciocínio indutivo, é necessário eliminar os efeitos da inteligência fluida e da inteligência geral nesse escore. Caso contrário, o escore de raciocínio indutivo apresenta a influência de duas variáveis latentes em sua estimativa, enviesando-a.

Não é por coincidência que uma série de pesquisadores da literatura psicométrica mundial começaram a realizar procedimentos para separar os efeitos de variáveis latentes mais amplas para estimar a fidedignidade de variáveis latentes mais específicas e vice-versa. A lógica do modelo de Carroll (1993, 2003) propagou uma espécie de *Zeitgeist* atual, indicando um foco relevante para a questão da fidedignidade. Ao transpor a questão da fidedignidade de escores múltiplos para o modelo do Enem, o problema se situa em compreender até que ponto um fator geral de desempenho escolar afeta os escores nos domínios de Matemática, Ciências da Natureza, Ciências Humanas e Linguagens. Afinal, ao controlar o efeito de um fator geral de desempenho escolar, os escores dos domínios mostram-se ainda confiáveis? Essa é a questão deste estudo. O presente estudo investiga, pois, a fidedignidade dos escores do Enem nos domínios, tendo controlado o efeito do fator geral de desempenho escolar.

Método

Participantes

Foram analisados os escores de 66.880 estudantes que participaram do Enem de 2011 e completaram especificamente os cadernos 120,

124, 125 e 129 (Inep, 2012a). A média de idade dos participantes era de 21,48 anos (*DP* = 7,12); sendo 53,3% do sexo feminino. Quanto à raça/cor, 50,5% eram 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.

Instrumento

A prova objetiva do Enem de 2011 foi composta por 180 itens, sendo dividida em quatro grupos de 45 itens que sustentam a medida para cada um dos quatro domínios: 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). Em cada domínio, há um conjunto de competências de área que, por sua vez, possuem um conjunto de habilidades específicas. A prova de 2011 teve a seguinte composição. Em LC, as nove competências de área foram aferidas por quatro a oito itens, e as habilidades por um até três itens. Em MT, as sete competências de área foram aferidas por grupos de quatro a oito itens, e as habilidades representadas por um ou dois itens, com exceção da habilidade 14, que não foi representada por nenhum item. Em CN, as oito competências de área possuíam de quatro a oito itens e as habilidades foram representadas por um ou dois itens, com exceção da habilidade 24 que possui três itens. Por fim, em CN, as seis competências de área continham de seis a 10 itens para sua aferição e as habilidades foram representadas por um ou dois itens, no máximo. Ao longo deste texto, as competências de área são representadas pela sigla do domínio seguida por um número (e.g., LC1, MT2, CN3, CN4), utilizando a mesma nomenclatura da Matriz de Referência do Exame (Inep, 2012b). A estrutura detalhada da prova de 2011 pode ser consultada no manual do usuário dos microdados do Enem de 2011 (Inep, 2012a).

A prova objetiva de 180 itens do Enem é realizada em dois encontros de 4 horas. No primeiro encontro o estudante responde a 90 itens, e no segundo encontro aos outros 90 itens, do total de 180 itens. Todos itens são de múltipla-escolha.

Procedimentos

Os dados utilizados neste estudo dizem respeito aos microdados do Enem de 2011 (Inep, 2012a). O *download*, a extração, a importação e o tratamento inicial dos dados foram realizados por meio do pacote *ENEM* (Golino, 2014), desenvolvido para o software *R* (R Core Team, 2018). O tratamento inicial envolveu a exclusão dos participantes ausentes nas provas e a correção dos vetores de resposta de acordo com o gabarito.

Análise de dados

Visando realizar a estimativa da fidedignidade dos quatro domínios do Enem, tomando como controle uma variável latente geral, conforme discutido na introdução, estimou-se os parâmetros de dois modelos concorrentes por meio de análise fatorial confirmatória (Beaujean, 2014; Kline, 2016), o Modelo dos Domínios Correlacionados e o Modelo Bifatorial. O primeiro modelo define a presença de quatro variáveis latentes correlacionadas, que representam os domínios de Matemática, Linguagens, Ciências da Natureza e Ciências Humanas. Cada uma dessas variáveis latentes explica unicamente os escores dos estudantes nas respectivas competências de área-alvo. Por exemplo, a variável latente de Matemática explica as competências de área em Matemática, a variável latente de Ciências Humanas explica as competências de área em Ciências Humanas, e assim por diante. O segundo modelo, denominado Bifatorial, apresenta as mesmas relações entre os domínios e suas competências de área, mas determina também a presença de um fator geral de desempenho escolar, assim como determina que todas variáveis latentes não se correlacionam, de maneira que os escores fatoriais dos domínios não sejam influenciados pelo fator geral. O controle do efeito do fator geral sobre os escores fatoriais dos quatro domínios permite verificar a fidedignidade real da aferição específica nos domínios.

A estimação dos modelos foi realizada utilizando-se o pacote *lavaan* (Rosseel et al., 2019) do R (R Core Team, 2018). Os modelos foram testados e comparados empregando-se o *weighted least*

squares mean adjusted estimator (WLSM). O uso do WLSM justifica-se dado que a natureza dos indicadores das variáveis latentes é categórica, pois esses são constituídos pelos escores totais nas competências de cada domínio. O ajuste dos modelos foi verificado por meio da análise da raiz do quadrado do erro de aproximação (RMSEA), do índice de ajuste comparativo (CFI) e do índice de qualidade do ajuste (GFI). Como critérios para um bom ajuste aos dados, indica-se que o RMSEA seja igual ou inferior a 0,06, e o CFI e o GFI sejam iguais ou superiores a 0,95 (Cangur & Ercan, 2015; Kline, 2016).

A análise da fidedignidade envolveu a estimação da fidedignidade composta e do índice *omega* (Gomes, Golino, & Peres, 2018; McDonald, 1999; Raykov, 2001; Revelle & Condon, 2018; Valentini et al., 2015). Uma vez que esses não são tão conhecidos, suas fórmulas serão apresentadas a seguir. A fidedignidade composta é calculada da seguinte forma:

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

Para o cálculo da fidedignidade composta, soma-se os betas (i.e., cargas fatoriais padronizadas) relacionados a uma determinada variável latente e as variáveis observáveis. Essa soma é elevada ao quadrado e essa operação é traduzida na seguinte parte da equação: $(\sum \text{betas})^2$. Por sua vez, soma-se os erros de cada variável observável. Os erros representam a variância não explicada da variável observável pela variável latente. Por exemplo, se uma variável latente possui um beta de 0,3 em relação a uma determinada variável observável, o erro será de 1 menos 0,3 elevado ao quadrado, indicando o valor de 0,91, ou seja, 91% da variância nessa variável observável não é explicada pela variável latente em questão e, por isso, é entendida como erro. Esse valor de 0,91 é somado aos 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 $(\sum \text{betas})^2 + \sum \text{erros}$. Esse resultado é usado para servir de divisor em relação aos próprios betas somados e elevados ao quadrado, o que define a fidedignidade composta. Como pôde ser observado, a fidedignidade composta inclui os betas e estima de forma correta a fidedignidade em modelos multidimensionais, como é o caso do Enem. Hair, Black, Babin e Anderson (2013) sugerem um ponto de corte de 0,70 para a fidedignidade composta.

Já o índice *omega* (Raykov, 2001) é calculado por meio da seguinte fórmula:

$$\omega_1 - \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2 Var(\psi)}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 Var(\psi) + \sum_{i=1}^{k} \theta_{ii} + 2\sum_{i<j} \theta_{ij}}$$

onde λ_i é a carga fatorial do item *i*, $Var(\psi)$ é a variância do fator, θ_{ii} é a variância do erro de medida do item *i*, e θ_{ij} é a covariância do erro de medida dos itens *i* e *j*. O índice *omega* foi calculado utilizando o pacote *semTools* (Jorgensen, Pornprasertmanit, Schoemann, & Rosseel, 2018) do R (R Core Team, 2018).

Resultados

Ambos os modelos apresentam adequado grau de ajuste aos dados, embora o Modelo Bifatorial tenha apresentado melhor ajuste (Tabela 1). Os índices de ajuste do Modelo dos Domínios Correlacionados são os seguintes: $\chi^2[399] = 28900{,}55$; CFI = 0,993; GFI = 0,998; RMSEA = 0,033. Já o Modelo Bifatorial apresentou os seguintes índices: $\chi^2[374] = 18005{,}49$; CFI = 0,995; GFI = 0,998; RMSEA = 0,027. A Tabela 1 também apresenta as médias das cargas fatoriais e dos erros padronizados. As Figuras 1 e 2 apresentam os dois modelos com os betas e correlações obtidos para cada competência.

Tabela 1 – Índices de ajuste dos modelos

Índice de ajuste	Domínios	Modelo dos Domínios Correlacionados	Modelo Bifatorial
χ^2	-	28900,55 (GL = 399; p < 0,000)	18005,49 (GL = 375; p < 0,000)
CFI	-	0,993	0,995
GFI	-	0,998	0,998
RMSEA (IC 95%)	-	0,033 (0,032-0,033)	0,027 (0,026-0,027)
CF (EP)	LC	0,66 (0,08)	0,25 (0,32)
	MT	0,69 (0,07)	0,39 (0,03)
	CN	0,55 (0,15)	0,16 (0,07)
	CH	0,69 (0,07)	0,18 (0,04)
	DEG	-	0,59 (0,11)

Legenda: χ^2 (Qui-Quadrado); GL (Graus de Liberdade); CFI (Índice de Ajuste Comparativo); GFI (Índice de Qualidade do Ajuste); RMSEA (Raiz do Quadrado do Erro de Aproximação); IC 95% = intervalo de confiança de 95% (limite inferior e limite superior); CF (média das cargas fatoriais); EP (média do erro padrão); LC (Linguagens, Códigos e suas Tecnologias); MT (Matemáticas e suas Tecnologias); CN (Ciências da Natureza e suas Tecnologias); CH (Ciências Humanas e suas Tecnologias); DEG (fator geral, denominado Desempenho Escolar Geral).

Figura 1 – Modelo dos Domínios Correlacionados do ENEM 2011, com betas e correlações. Os círculos correspondem às variáveis latentes (i.e., domínios), enquanto os quadrados correspondem aos seus indicadores (i.e., competências de área). Legenda: LC (Linguagens, Códigos e suas Tecnologias); MT (Matemáticas e suas Tecnologias); CN (Ciências da Natureza e suas Tecnologias); CH (Ciências Humanas e suas Tecnologias).

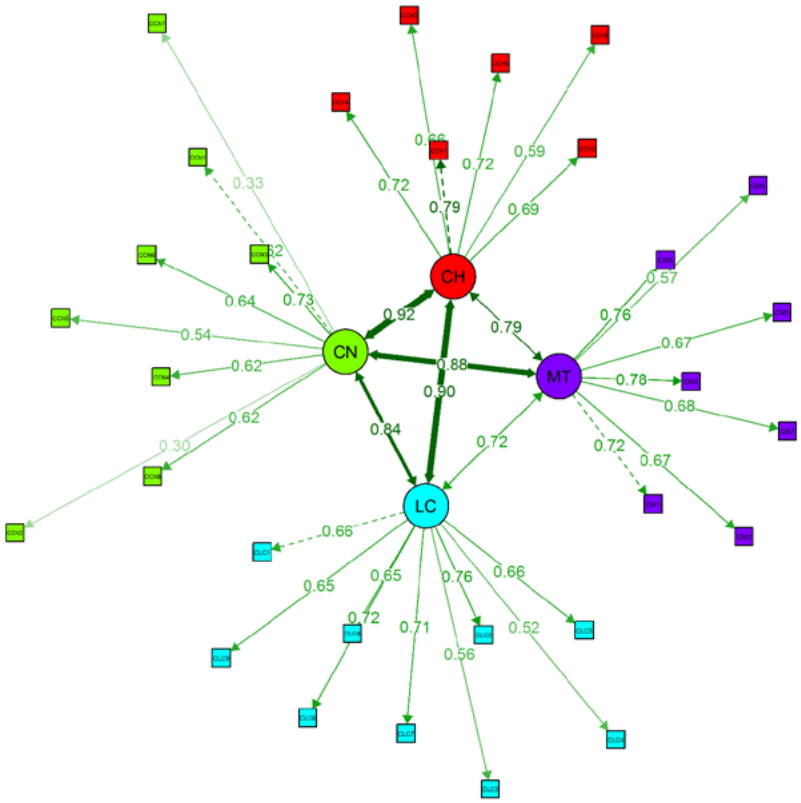
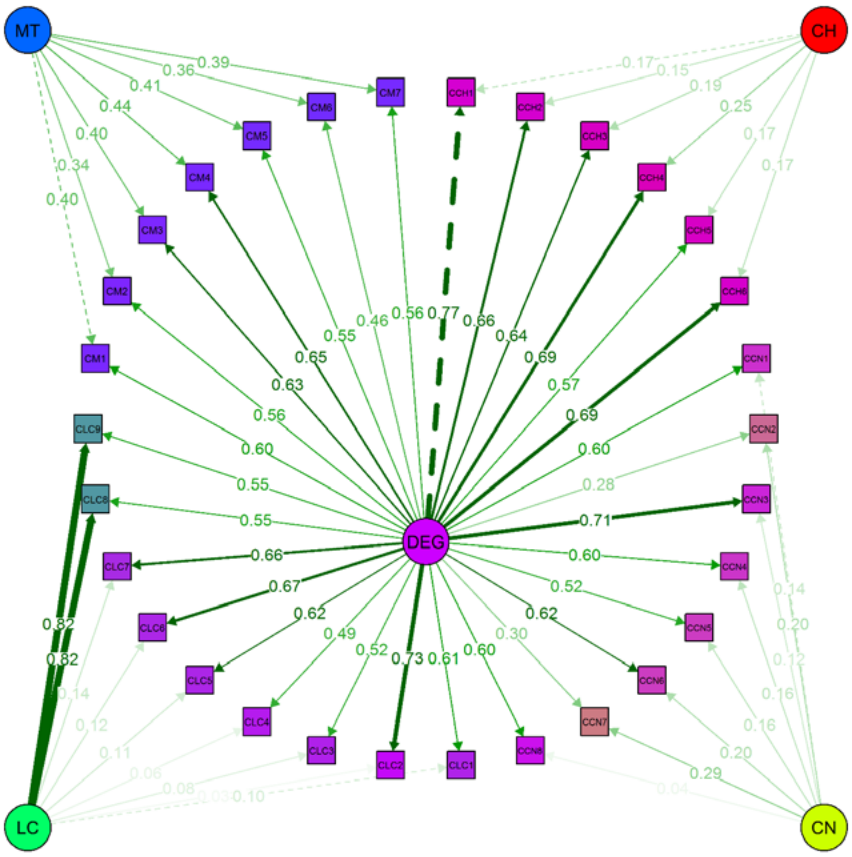


Figura 2 – Modelo Bifatorial do ENEM 2011, com betas e correlações. Os círculos correspondem às variáveis latentes (i.e., domínios), enquanto os quadrados correspondem aos seus indicadores (i.e., competências de área). Legenda: DEG (fator geral, denominado Desempenho Escolar Geral); LC (Linguagens, Códigos e suas Tecnologias); MT (Matemáticas e suas Tecnologias); CN (Ciências da Natureza e suas Tecnologias); CH (Ciências Humanas e suas Tecnologias).



A Tabela 2 apresenta os valores obtidos para os coeficientes de fidedignidade composta e *omega*. Para o Modelo dos Domínios Correlacionados, a fidedignidade composta apresentou os seguintes valores: Linguagens (0,872), Matemática (0,867), Ciências da Natureza (0,781) e Ciências Humanas (0,849). No que se refere ao coeficiente *omega*, no Modelo dos Domínios Correlacionados os seguintes valores foram encontrados: Linguagens (0,877), Matemática (0,872), Ciências da Natureza (0,797) e Ciências Humanas (0,853). De acordo com as interpretações sugeridas por Hogan (2013), verifica-se que em todos os domínios os valores são considerados bons (i.e., entre 0,80 e 0,89), com exceção de CN. De acordo com Hogan, coeficientes de fidedignidade entre 0,70 e 0,79 requerem informações suplementares que sustentem o uso dos escores do teste.

Por sua vez, para o Modelo Bifatorial, a fidedignidade composta apresentou os seguintes valores: Linguagens (0,407), Matemática (0,558), Ciências da Natureza (0,184) e Ciências Humanas (0,174). O fator de desempenho escolar geral apresentou uma fidedignidade composta de 0,941. Já o coeficiente *omega* apresentou os seguintes valores: Linguagens (0,477), Matemática (0,683), Ciências da Natureza (0,244), Ciências Humanas (0,292) e o fator de desempenho escolar geral (0,952). Verifica-se que a introdução do fator geral ao modelo interferiu negativamente na fidedignidade das variáveis latentes correspondentes aos domínios do Enem. Apenas o fator geral apresentou valor satisfatório, considerado excelente de acordo com a interpretação sugerida por Hogan (2013). Ainda de acordo com essa interpretação, o coeficiente *omega* do domínio de matemática pode ser considerado para propósitos de pesquisa.

Tabela 2 – *Fidedignidade composta e coeficiente Omega*

Domínios	Modelo dos Domínios Correlacionados		Modelo Bifatorial	
	Fidedignidade composta	***Omega***	**Fidedignidade composta**	***Omega***
MT	0,867	0,872	0,558	0,683
CN	0,781	0,797	0,184	0,244
LC	0,872	0,877	0,407	0,477
CH	0,849	0,853	0,174	0,292
DEG	-	-	0,941	0,952

Legenda: DEG (fator geral, denominado Desempenho Escolar Geral); LC (Linguagens, Códigos e suas Tecnologias); MT (Matemáticas e suas Tecnologias); CN (Ciências da Natureza e suas Tecnologias); CH (Ciências Humanas e suas Tecnologias).

Conclusão

Os resultados indicaram que a fidedignidade dos escores do Modelo dos Domínios Correlacionados pode ser considerada boa no geral, variando entre 0,781 (CN) e 0,872 (LC) no caso da fidedignidade composta, e entre 0,797 (CN) e 0,877 (LC) no caso do coeficiente *omega*. No entanto, no caso do Modelo Bifatorial, os coeficientes de fidedignidade referentes aos escores não podem ser considerados satisfatórios, variando entre 0,174 (CH) e 0,558 (MT) no caso da fidedignidade composta, e entre 0,244 (CN) e 0,683 (MT) no caso do *omega*. Apesar disso, o coeficiente de fidedignidade composta do fator de desempenho escolar geral foi de 0,941 e o *omega* de 0,952. Ou seja, para essa variável latente, os valores observados atendem proximamente ao critério de 0,95 indicado por Hogan (2013) para testes *high stake*.

Os resultados indicam que as competências de área são consideravelmente influenciadas pelo fator geral escolar, de modo que as medidas de domínio, geradas pelos escores nas competências de área, apresentaram uma fidedignidade baixa. Essa característica, por sua vez, pode não ser uma característica restrita da prova de 2011 do Enem, mas situar-se como uma propriedade presente também nas outras provas. Isso porque as provas do Enem têm como foco possuir itens que enfoquem o caráter interdisciplinar e a resolução de problemas. Um dos elementos-chaves nos itens do Enem envolve a capacidade do estudante em ler atentamente as instruções e interpretar as dicas presentes nos próprios itens. Outra característica saliente nos itens é a presença de diferentes estímulos e conteúdos envolvidos. Por exemplo, nos itens de Matemática a habilidade de leitura e de interpretação dos enunciados é bastante relevante. Essa diversidade de habilidades envolvidas em todos os domínios pode ativar intensamente o fator geral de desempenho escolar, na medida em que nenhuma habilidade escolar muito específica é fortemente mobilizada nas provas.

Do ponto de vista teórico, cabe ainda argumentar que, embora esses resultados referentes à fidedignidade dos escores da prova de 2011 possam parecer exclusivamente negativos, há uma perspectiva positiva para sua interpretação. Os resultados indicam que o Enem pode ser um instrumento de forte aferição do desempenho escolar geral do aluno. Do ponto de vista desenvolvimental, essa habilidade é muito importante, bem mais, inclusive, do que os próprios domínios. Se os domínios representam uma antiga tradição escolar de compreender os conteúdos e organizar o ensino e o espaço escolar, o fator escolar geral representa teoricamente a capacidade dos estudantes de estruturar de forma verdadeiramente interdisciplinar as diferentes disciplinas veiculadas ao longo de sua trajetória discente.

Essa interpretação dos resultados vai ao encontro das formulações feitas por Carroll (1993, 2003) em relação aos três estratos de habilidades cognitivas por ele identificados para explicar o desempenho em testes de inteligência. No Enem, de acordo com os modelos aqui explorados, temos como primeiro extrato as competências de área, que descrevem processos cognitivos específicos. Por exemplo, a Competência de Área 3 (CM3) do domínio Matemática requer "construir noções de grandezas e medidas para a compreensão da realidade e a solução de problemas do cotidiano" (Inep, 2012b, p. 5). No segundo extrato, por sua vez, temos os quatro domínios do Enem (i.e., MT, LC, CN e CH), que envolvem processos cognitivos mais amplos que as competências de área, arbitrariamente separados em quatro grandes áreas das ciências. Como terceiro extrato, que envolve uma habilidade cognitiva geral nas formulações de Carrol, no caso do Enem temos o fator geral de desempenho escolar. Assim, os escores dos participantes do Enem precisam ser compreendidos como resultantes dos efeitos desses três extratos.

Retomando a questão do estudo, ao controlar o efeito de um fator geral de desempenho escolar, os escores dos domínios mostram-se ainda confiáveis? Em resposta, o presente estudo mostra a importância de se controlar o efeito do fator geral sobre os escores gerados nos domínios do Enem. Sem esse controle, é possível que a fidedignidade dos escores seja superestimada. No entanto, esse estudo não permite concluir que as medidas dos quatro domínios apresentam baixa fidedignidade, em função da análise ter se limitado aos escores agregados das competências de área. Como as competências reduzem o número de variáveis observáveis a 30, pois são 30 as competências de área, a fidedignidade composta e o *omega* podem ter se mostrado reduzidos em função do baixo número de variáveis observáveis. Sabe-se que quanto maior o número de variáveis observáveis, maior tende a ser a fidedignidade da medida, seja ela a fidedignidade estimada pela fidedignidade composta ou pelo *omega*. Assim, uma análise precisa da fidedignidade dos domínios do Enem obrigatoriamente demanda a investigação da fidedignidade dos escores nos quatro domínios, tendo como variáveis observáveis os 180 itens, e não os escores nas 30 competências. Concomitante a isso, a fidedignidade composta deve ser calculada a partir de modelos provenientes da análise fatorial de itens. Essa análise deve ser realizada em estudo posterior.

Referências

Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*. Routledge. (DOI INEXISTENTE)

Cangur, S., & Ercan, I. (2015). Comparison of model fit indices used in structural equation modeling under multivariate normality. *Journal of Modern Applied Statistical Methods*, *14*(1), 152-167. 10.22237/jmasm/1288584240

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. London, UK: Cambridge University Press. 10.1017/CBO9780511571312

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports g and about ten broad factors. In H. Nyborg (Ed.), *The scientific study of general intelligence: tribute to Arthur R. Jensen* (pp. 1-20). Oxford, UK: Pergamon Press. 10.1016/B978-008043793-4/50036-2

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. (DOI INEXISTENTE)

Golino, H. F., Gomes, C. M. A., Amantes, A., & Coelho, G. (Eds.). (2015). *Psicometria contemporânea. Compreendendo os modelos Rasch.* Casa do Psicólogo. (DOI INEXISTENTE)

Gomes, C. M. A. (2010). Avaliando a avaliação escolar: notas escolares e inteligência fluida. *Psicologia em Estudo, 15*(4), 841-849. Recuperado de http://www.redalyc.org/articulo.oa?id=287123084020

Gomes, C. M. A., & Borges, O. (2008). Limite da validade de um instrumento de avaliação docente. *Avaliação Psicológica*, *7*(3), 391-401. Recuperado de http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677-04712008000300011&lng=pt&tlng=pt.

Gomes, C. M. A. & Borges, O. N. (2009). O ENEM é uma avaliação educacional construtivista? Um estudo de validade de construto. *Estudos em Avaliação Educacional, 20*(42), 73-88. 10.18222/eae204220092060

Gomes, C. M. A., Golino, H. F., & Peres. (2016). Investigando a validade estrutural das competências do ENEM: quatro domínios correlacionados ou um modelo bifatorial. *Boletim na Medida (INEP-Ministério da Educação), 5*(10), 33-30. Recuperado de http://portal.inep.gov.br/documents/186968/494037/BOLETIM+NA+MEDIDA+-+N%C2%BA+10/4b8e3d73-d95d-4815-866c-ac2298dff0bd?version=1.1

Gomes, C. M. A., Golino, H. F., & Peres, A. J. S. (2018). Análise da fidedignidade composta dos escores do ENEM por meio da análise fatorial de itens. *European Journal of Education Studies, 5*(8), 331-344. 10.5281/zenodo.2527904

Gomes, C. M. A., & Jelihovschi, E. (2019). Presenting the regression tree method and its application in a large-scale educational dataset. *International Journal of Research & Method in Education.* 10.1080/1743727X.2019.1654992

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

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

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2012a). *Microdados do ENEM – 2011. Exame Nacional do Ensino Médio: Manual do Usuário.* Recuperado de http://portal.inep.gov.br/web/guest/microdados (DOI INEXISTENTE)

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2012b). *Matriz de Referência do ENEM.* Recuperado de http://download.inep.gov.br/educacao_basica/enem/downloads/2012/matriz_referencia_enem.pdf (DOI INEXISTENTE)

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2018). *Edital nº 16, de 20 de Março de 2018. Exame Nacional Do Ensino Médio - Enem 2018.* Recuperado de http://download.inep.gov.br/educacao_basica/enem/edital/2018/edital_enem_2018.pdf (DOI INEXISTENTE)

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2018). semTools: Useful tools for structural equation modeling (version 0.5-1). [Computer software]. Recuperado de https://CRAN.R-project.org/package=semTools

Kline, R. B. (2016). *Principles and practice of Structural Equation Modeling* (4th ed.). New York, USA: The Guilford Press. (DOI INEXISTENTE)

McDonald, R. P. (1999). *Test theory: A unified treatment*. L. Erlbaum Associates (DOI INEXISTENTE)

Pires, A. A. M., & Gomes, C. M. A. (2017). Three mistaken procedures in the elaboration of school exams: explicitness and discussion. *PONTE International Scientific Researches Journal, 73*(3), 1-14. 10.21506/j.ponte.2017.3.1

Pires, A. A. M., & Gomes, C. M. A. (2018). Proposing a method to create metacognitive school exams. *European Journal of Education Studies, 5(8),* 119-142. 10.5281/zenodo.2313538

R Core Team (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Recuperado de http://www.R-project.org/ (DOI INEXISTENTE)

Raykov, T. (2001). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology, 54*, 315-323. 10.1348/000711001159582

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47* (5), 667-696. 10.1080/00273171.2012.715555

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. (DOI INEXISTENTE)

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema, 26* (1), 108-116. 10.7334/psicothema2013.260

Rosseel, Y., Jorgensen, T. D., Oberski, D., Vanbrabant, J. B. L., Savalei, V., Hallquist, E. M., M., Rhemtulla, M., Katsikatsou, M., Barendse, M., Scharf, F. (2019). lavaan: Latent Variable Analysis (version 0.6-5). [Computer software]. Recuperado de https://cran.r-project.org/web/packages/lavaan/index.html (DOI INEXISTENTE)

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. 10.15348/1980-6906/psicologia.v17n2p123-139

Cristiano Mauro Assis Gomes

Doutor em Educação pela Universidade Federal de Minas Gerais (UFMG, Belo Horizonte, MG, Brasil), professor do Departamento de Psicologia da Universidade Federal de Minas Gerais em Belo Horizonte, MG, Brasil.

Hudson Fernandes Golino

Doutor em Neurociências pela Universidade Federal de Minas Gerais (UFMG, Belo Horizonte, MG, Brasil), professor do Departamento de Psicologia da University of Virginia em Charlottesville, VA, Estados Unidos da América.

Alexandre José de Souza Peres

Doutor em Psicologia Social, do Trabalho e das Organizações pela Universidade de Brasília (UNB, Brasília, DF, Brasil). Professor do Câmpus de Paranaíba da Universidade Federal de Mato Grosso do Sul em Paranaíba, MS, Brasil.

Endereço para correspondência

Cristiano Mauro Assis Gomes

Universidade Federal de Minas Gerais

Av. Antônio Carlos, 6627, gabinete 4036

Pampulha, 31270-901

Belo Horizonte, MG, Brasil