Psicologia: Teoria e Prática, 23(1), 1-22. São Paulo, SP, 2021.
ISSN 1516-3687 (impresso), ISSN 1980-6906 (*on-line*). **Sistema de avaliação**: às cegas por pares (*double blind review*).
Universidade Presbiteriana Mackenzie.

Avaliação Psicológica

Criando uma medida verdadeira para o Enem: Uma análise pelo Modelo Rasch

Hudson F. Golino[1]
https://orcid.org/0000-0002-1601-1447

Cristiano Mauro A. Gomes[2]
https://orcid.org/0000-0003-3939-5807

Alexandre José de S. Peres[3]
https://orcid.org/0000-0002-3472-6120

Para citar este artigo: Golino, H. F., Gomes, C. M. A., & Peres, A. J. S. (2021). Criando uma medida verdadeira para o Enem: Uma análise pelo Modelo Rasch. *Psicologia: Teoria e Prática*, 23(1), 1–22.

Submissão: 02/07/2019
Aceite: 23/06/2020

1 Universidade da Virginia (UVA), Charlottesville, VA, Estados Unidos.

2 Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, MG, Brasil.

3 Universidade Federal de Mato Grosso do Sul (UFMS), Paranaíba, MS, Brasil.

Resumo

Nos anos 1930, um grupo de cientistas argumentou que a concatenação empírica de elementos observáveis não seria possível nas Ciências Humanas e Sociais e por isso era inviável obter medidas verdadeiras nesses campos do conhecimento científico. Para lidar com esse problema, foram propostas teorias matemáticas nas quais a concatenação empírica não seria necessária, como a Teoria de Medidas Aditivas Conjuntas (TMAC). No mesmo período, George Rasch desenvolveu o modelo logístico simples para dados dicotômicos, uma operacionalização probabilística da TMAC que viabiliza a análise empírica de pressupostos da medida verdadeira. Em nosso estudo, investigamos o desenvolvimento de uma medida verdadeira para o Exame Nacional do Ensino Médio (Enem), aplicando o modelo logístico simples em dados referentes à *performance* dos participantes da edição de 2011. Os resultados indicaram um ajuste adequado do modelo, apontando para a viabilidade da construção de uma medida verdadeira para o Enem. Implicações são discutidas.

Palavras-chave: Exame Nacional do Ensino Médio (Enem); Teoria de Medidas Aditivas Conjuntas; modelo de Rasch; Teoria de Resposta ao Item; avaliação educacional.

CREATING A TRUE MEASUREMENT FOR ENEM: AN ANALYSIS USING THE RASCH MODEL

Abstract

In the 1930s, a group of scientists argued that empirical concatenation of observable elements was not possible in Human and Social Sciences and, therefore, was not feasible to obtain true measurements similar to those found in Physics. To deal with this issue, mathematical theories that doesn't require concatenation have been proposed in the 1960s, such as the Additive Conjoint Measurement Theory (ACMT). In the same decade, George Rasch developed the simple logistic model for dichotomous data, a probabilistic operationalization of ACMT. This study investigates the possibility of developing a fundamental measurement for the National Exam of Upper Secondary Education (ENEM) applying Rasch's model to students' performance in the 2011 ENEM exam. The results indicated an adequate model fit, which point out for the viability of a fundamental measure using ENEM's data. Implications are discussed.

Keywords: National Exam of Upper Secondary Education (ENEM); Theory of Additive Conjoint Measurement; Rasch model; Item Response Theory; educational assessment.

CREANDO UNA VERDADERA MEDIDA PARA ENEM: UN ANÁLISIS POR EL MODELO RASCH

Resumen

En los años 1930, un grupo de científicos argumentó que la concatenación empírica de elementos observables no sería posible en Ciencias Humanas y Sociales y por consiguiente sería inviable obtener medidas verdaderas similares a las de Física. Para abordar este problema, a partir de los años 1960 se proponían teorías en las cuales la concatenación empírica no es necesaria, como la Teoría de Medidas Aditivas Conjuntas (TMAC). Al mismo período, George Rasch desarrolló el modelo logístico simple para datos dicotómicos, una operacionalización probabilística de la TMAC. Este estudio investigó la posibilidad de desarrollar una medida verdadera para el Examen Nacional de la Secundaria Superior (ENEM), aplicando el modelo logístico simple en los datos referentes a la performance de los participantes en la prueba de 2011. Los resultados indicaron adecuado ajuste del modelo, así como la viabilidad de una medida verdadera para el ENEM. Implicaciones son discutidas.

Palabras clave: Examen Nacional de la Secundaria Superior (ENEM); Teoría de Medidas Aditivas Conjuntas; modelo de Rasch; Teoría de Respuesta al Ítem; evaluación educativa.

1. Introdução

Na década de 1930, um grupo de pesquisadores da Física e da Psicologia reuniu-se na Associação Britânica para o Avanço da Ciência para discutir a viabilidade da mensuração em Psicologia, Educação e áreas correlatas (Borsboom, 2005). Não houve consenso, mas a maioria acompanhou os apontamentos de Campbell (1920) de que seria impossível desenvolver qualquer tipo de medida nas Ciências Sociais e Humanas em geral, pois os objetos de estudo dessas áreas não permitiam concatenar objetos para criar sistemas de comparação de quantidades. Na época, a medida era definida e operacionalizada por meio da abordagem do representacionalismo clássico (Borsboom, 2005). Nela, a concatenação era considerada peça fundamental e obrigatória para a geração de uma medida, pois, por meio dela, o sistema empírico (de relações observadas na natureza) podia ser mapeado em um sistema representacional (de números e operações matemáticas de comparação; ver Golino & Gomes, 2015), gerando uma medida que representava as características do objeto de forma correta ou verdadeira.

Após um longo período, o trabalho seminal de Krantz, Suppes, Luce e Tversky (1971) mostrou que a concatenação não seria condição obrigatória para a ocorrência de um mapeamento adequado entre o sistema empírico (isto é, objetos) e o sistema representacional (isto é, números) e, por consequência, para a geração de uma medida verdadeira – também chamada de medida fundamental. Esses autores fundaram uma nova área, denominada representacionalismo contemporâneo, em que axiomatizaram a teoria da medida e definiram matematicamente uma série de propriedades fundamentais que resultam em medidas numéricas adequadas, tanto para a Física, a Geometria e outras áreas das Ciências Exatas quanto para a Educação, a Psicologia e áreas correlatas. Os autores contrapõem o representacionalismo clássico, afirmando que é errado pensar que apenas um único sistema formal de relações leva à medida verdadeira. Eles mostram que a própria Física trabalha com a mensuração de atributos que não são passíveis de operações de concatenação empírica, como a temperatura, por exemplo.

Para obter uma medida verdadeira sem a necessidade da concatenação, Krantz et al. (1971) propuseram a Teoria de Medidas Aditivas Conjuntas (TMAC). Nessa abordagem, as regras a serem seguidas no processo de mapeamento do sistema relacional no sistema numérico são estritas e devem satisfazer quatro axiomas (Borsboom, 2005; Golino & Gomes, 2015). Para facilitar a compreensão, os axiomas serão apresentados por meio de um exemplo. Suponha que se esteja interessado em medir um objeto ou um atributo, como a habilidade em Matemática, e que esse atributo seja estudado por meio de duas dimensões (isto é, variáveis independentes) conjuntas: a habilidade matemática das pessoas e a dificuldade dos itens para avaliar a habilidade matemática. A realização conjunta dessas dimensões (isto é, o encontro das pessoas com os itens) gera uma terceira variável, essa dependente, que é a resposta das pessoas. Quando há um mapeamento adequado do sistema de relações qualitativas verificado na variável dependente em um sistema numérico que representa essas relações, deve-se produzir quatro consequências, que representam os axiomas da TMAC.

A primeira consequência (Axioma 1 da TMAC) é que o valor de uma das dimensões, a habilidade, pode ser escolhido sem afetar o valor da outra dimensão, a dificuldade dos itens, indicando uma separação entre o que está sendo medido e o objeto de medida, condição necessária para uma medida dos atributos (Thurstone, 1931). Nesse sentido, a habilidade de uma pessoa não afeta a estimativa de dificuldade de um item, nem a dificuldade de um item afeta a estimativa da habilidade de uma pessoa.

A segunda consequência (Axioma 2), que oriunda diretamente da primeira, é o ordenamento independente da habilidade e da dificuldade ao longo da medida construída (isto é, da habilidade em Matemática). Em outras palavras, pessoas com mais habilidade terão uma posição maior na escala de medida do que pessoas com menos habilidade, independentemente dos itens que forem utilizados para aferir essa habilidade. De forma análoga, itens mais difíceis terão uma posição maior na escala de medida do que itens mais fáceis, independentemente de quais pessoas responderam a esses itens e os acertaram.

A terceira consequência (Axioma 3) é a de que um aumento quantitativo na medida produzida resulta em efeitos específicos na habilidade e na dificuldade, mas de forma independente uma da outra. Por fim, a quarta consequência (Axioma 4) implica que as habilidades das pessoas são comparáveis, de modo que a diferença entre os escores das pessoas possui um significado que reflete diferenças reais nas habilidades. Da mesma forma, as dificuldades dos itens são comparáveis, de modo que a diferença nos escores dos itens reflete diferenças reais de dificuldade entre eles.

A despeito de os axiomas de Krantz et al. (1971) serem uma alternativa ao modelo clássico, eles não seriam efetivos sem a presença de um tratamento estatístico capaz de verificar se as quantificações produzidas nas áreas de Ciências Humanas e Sociais atendem a esses axiomas e podem ser avaliadas como medidas verdadeiras (ver Bond & Fox, 2015; Golino & Gomes, 2015). Os modelos psicométricos de George Rasch (1960) eliminaram esse problema ao definir funções que possibilitam o mapeamento das relações qualitativas em um sistema representacional numérico que obedece aos axiomas da medida de Krantz et al. (1971). Em sua racionalidade, os modelos Rasch verificam estatisticamente se a estrutura dos dados oriundos de quantificações provenientes de instrumentos de medida (por exemplo, provas educacionais, testes psicológicos, entre outros) ajustam-se às relações do tipo aditivas conjuntas que satisfazem os quatro axiomas da medida.

Quando não há ajuste dos dados aos modelos Rasch, pode-se concluir que a quantificação não reflete uma estrutura aditiva conjunta e, por consequência, uma medida verdadeira. Do ponto de vista metodológico, os modelos Rasch buscam anomalias nas quantificações que as distanciem de um critério operacional, matematicamente bem definido, ao qual as quantificações deveriam se ajustar para sustentarem uma medida verdadeira. Não por acaso, Andrich (2004, p. 12) afirma que "identificar anomalias substantivas a partir da análise de desajuste, resistindo à modificação do modelo, [e] coletando novos dados guiados pelo modelo é consistente com o papel da medida nas ciências físicas como enunciado por Kuhn".

Considerando que os modelos Rasch são cruciais para a efetivação dos axiomas da TMAC, sua racionalidade será demonstrada neste artigo. No entanto, apenas apresentaremos o modelo dicotômico, por ser o mais simples e por ser suficiente para essa demonstração. Também chamado de modelo logístico simples (MLS), esse modelo define que a resposta X_{pi}, que surge do encontro da pessoa *p* com o item *i*, depende da habilidade β da pessoa e da dificuldade δ do item, expressada em termos probabilísticos. A probabilidade de a pessoa acertar um determinado item varia de acordo com a sua habilidade β. Dessa forma, se β_p for igual à δ_i, estima-se que a pessoa tem 50% de chance de acertar o item. Caso β_p seja menor do que δ_i, espera-se que a pessoa tenha menos que 50% de chance de acerto. No entanto, se β_p for maior do que δ_i, espera-se que a pessoa tenha mais que 50% de chance de responder corretamente. A relação entre habilidade e dificuldade é representada pela seguinte relação matemática genérica para respostas dicotômicas:

$$P\left\{X_{pi} = x_{pi}\right\} = \frac{e^{x_{pi}(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}} \qquad (1)$$

Entre as várias propriedades do modelo Rasch para dados dicotômicos, a invariância pode ser apontada como uma das mais importantes. Essa propriedade garante que os parâmetros do objeto medido e do instrumento de medida são separáveis, ou seja, comparações da habilidade das pessoas independem da dificuldade dos itens e vice-versa. Essa é uma propriedade do modelo matemático e não dos dados empíricos em si (Wright & Stone, 1999). Em um par de itens, a probabilidade de uma pessoa acertar o primeiro e errar o segundo, dado que ela acerta apenas um dos dois, depende única e exclusivamente da dificuldade desses itens. Essa propriedade pode ser verificada a seguir. Suponha que uma pessoa (*p*) responda a dois itens dicotômicos: item 1 e item 2. Os possíveis resultados são: 1) ela erra ambos os itens; 2) ela erra o primeiro e acerta o segundo; 3) ela acerta o primeiro e erra o segundo; 4) ela acerta ambos os itens. Considere, agora, que a pessoa *p* acerte o primeiro item e erre o segundo. Essa probabilidade é calculada como:

Psicologia: Teoria e Prática, *23*(1), 1-22, São Paulo, ePTPPA12625, 2021.
ISSN 1980-6909 (*on-line*)

$$P\{(x_{p1} = 1, x_{p2} = 0) \mid (x_{p1} = 1, x_{p2} = 0) \vee (x_{p1} = 0, x_{p2} = 1)\}$$
$$= \frac{\frac{e^{(\beta_p - \delta_1)}}{1 + e^{(\beta_p - \delta_1)}} \frac{1}{1 + e^{(\beta_p - \delta_2)}}}{\left(\frac{e^{(\beta_p - \delta_1)}}{1 + e^{(\beta_p - \delta_1)}} \frac{1}{1 + e^{(\beta_p - \delta_2)}}\right) + \left(\frac{1}{1 + e^{(\beta_p - \delta_1)}} \frac{e^{(\beta_p - \delta_2)}}{1 + e^{(\beta_p - \delta_2)}}\right)}$$

Apesar de a expressão de probabilidade acima ser grande, e parecer muito difícil de entender, ela é relativamente simples. O numerador é a probabilidade conjunta de a pessoa acertar o primeiro item e errar o segundo. O denominador é a probabilidade conjunta de a pessoa acertar o primeiro e errar o segundo item ou errar o primeiro e acertar o segundo. Vamos continuar desenvolvendo a equação de probabilidade:

$$\frac{e^{(\beta_p - \delta_1)} \frac{1}{1 + e^{(\beta_p - \delta_1)}} \frac{1}{1 + e^{(\beta_p - \delta_2)}}}{\left(e^{(\beta_p - \delta_1)} \frac{1}{1 + e^{(\beta_p - \delta_1)}} \frac{1}{1 + e^{(\beta_p - \delta_2)}}\right) + \left(\frac{1}{1 + e^{(\beta_p - \delta_1)}} \frac{1}{1 + e^{(\beta_p - \delta_2)}} e^{(\beta_p - \delta_2)}\right)}$$

Agora vamos isolar o produto da probabilidade de errar cada um dos itens no denominador da equação:

$$\frac{e^{(\beta_p - \delta_1)} \frac{1}{1 + e^{(\beta_p - \delta_1)}} \frac{1}{1 + e^{(\beta_p - \delta_2)}}}{\left(\frac{1}{1 + e^{(\beta_p - \delta_1)}} \frac{1}{1 + e^{(\beta_p - \delta_2)}}\right) e^{(\beta_p - \delta_1)} + e^{(\beta_p - \delta_2)}}$$

Podemos, agora, eliminar o produto da probabilidade de errar cada um dos itens, cancelando essa probabilidade presente no numerador com a probabilidade presente no denominador:

$$\frac{e^{(\beta_p - \delta_1)} \frac{1}{1 + e^{(\beta_p - \delta_1)}} \frac{1}{1 + e^{(\beta_p - \delta_2)}}}{\left(\frac{1}{1 + e^{(\beta_p - \delta_1)}} \frac{1}{1 + e^{(\beta_p - \delta_2)}}\right) e^{(\beta_p - \delta_1)} + e^{(\beta_p - \delta_2)}} = \frac{e^{(\beta_p - \delta_1)}}{e^{(\beta_p - \delta_1)} + e^{(\beta_p - \delta_2)}} =$$

No próximo passo, isolamos $e^{\beta p}$ no numerador e no denominador:

$$\frac{e^{\beta_p} e^{-\delta_1}}{e^{\beta_p}(e^{-\delta_1} + e^{-\delta_2})}$$

Por último, cancelamos $e^{\beta p}$ do numerador com o denominador:

$$\frac{e^{\beta_p} e^{-\delta_1}}{e^{\beta_p}(e^{-\delta_1} + e^{-\delta_2})} = \frac{e^{-\delta_1}}{e^{-\delta_1} + e^{-\delta_2}}$$

Assim, eliminamos o parâmetro da pessoa da equação. Em outras palavras, a probabilidade de uma pessoa *p* responder a um item 1 corretamente e a um item 2 incorretamente, dado que ela acerta um ou outro item apenas, é dado por uma relação envolvendo exclusivamente a dificuldade de ambos os itens (equação da suficiência dos itens). Ou seja:

$$P\{(x_{p1} = 1, x_{p2} = 0) \mid (x_{p1} = 1, x_{p2} = 0) \vee (x_{p1} = 0, x_{p2} = 1)\} = \frac{e^{-\delta_1}}{e^{-\delta_1} + e^{-\delta_2}} \quad (2)$$

De forma semelhante, tomando-se que duas pessoas respondam a um item *i*, a probabilidade de a primeira acertar esse item e de a segunda pessoa errá-lo, dado que apenas uma das duas acerta o item, depende única e exclusivamente da habilidade dessas pessoas. Essa propriedade é expressa da seguinte maneira:

$$P\{(x_{1i} = 1, x_{2i} = 0) \,\big|\, (x_{1i} = 1, x_{2i} = 0)\ OU\ (x_{1i} = 0, x_{2i} = 1)\} = \frac{e^{\beta_1}}{e^{\beta_1} + e^{\beta_2}} \quad (3)$$

A expressão de invariância dos parâmetros no modelo dicotômico de Rasch satisfaz um dos principais axiomas da Teoria da Medida Aditiva Conjunta, que é a da relação duplamente independente entre os fatores (no caso, habilidade e dificuldade). Ela é uma verificação matemática em que o modelo assume duas condições. A primeira define que o valor de β pode ser escolhido sem afetar o valor de δ (independência de realização dos componentes). A segunda define que o componente β e o componente δ possuem efeitos independentes no atributo a ser medido (no caso, uma variável latente). Dessa forma, o modelo dicotômico Rasch sustenta, matematicamente, o ordenamento independente de β e δ, ao longo da variável latente, satisfazendo o Axioma 2 da TMAC.

Além de satisfazer as duas condições do Axioma 2, a expressão de invariância dos parâmetros tem como consequência, também, que o aumento na variável latente produz um efeito específico de aumento na habilidade β e na dificuldade δ, mas de maneira independente uma da outra. Por consequência, o Axioma 3 da TMAC é satisfeito (duplo cancelamento). Por último, como a comparação da habilidade de duas pessoas β_1 e β_2 depende da relação entre as habilidades dessas pessoas, então os valores de β são comparáveis. Da mesma forma, como a comparação da dificuldade de dois itens δ_1 e δ_2 depende da relação entre as dificuldades desses itens, então os valores de δ também são comparáveis. Satisfaz-se, assim, o Axioma 4 da TMAC (Axioma de Arquimedes). Por último, se os dados de uma quantificação se ajustam ao modelo dicotômico de Rasch, então se conclui que a ordem entre as relações é do tipo fraca, satisfazendo o Axioma 1 da TMAC. Se a ordem das relações não é fraca, os dados não se ajustam ao modelo, e conclui-se que a quantificação analisada não sustenta uma medida verdadeira.

Acrescentamos que o modelo dicotômico de Rasch (1960) e os modelos dele derivados são as únicas funções probabilísticas existentes até o presente momento que mapeiam as relações qualitativas encontradas em estruturas aditivas conjuntas em um sistema representacional numérico, de forma que sejam satisfeitos todos os quatro axiomas da TMAC. Por consequência, salientamos que, para além dos modelos Rasch, nenhum outro modelo da Teoria de Resposta ao Item (TRI) ou nenhum modelo proveniente de outras metodologias permite esse tipo de análise. Entre os defensores dos modelos de dois e três parâmetros da TRI, há uma defesa de que os modelos Rasch são apenas uma versão simplificada dos modelos com mais parâmetros, uma posição epistemológica que vai contra os argumentos apresentados na literatura internacional de medida, explorada em detalhes por Andrich (2004). Ao adicionar parâmetros, perde-se um elemento fundamental que é a suficiência do escore total para a estimativa do parâmetro de habilidade das pessoas. Esse é o ponto central que faz com que os modelos de Rasch obtenham uma equação de suficiência dos itens que não possui o parâmetro de habilidade, possibilitando a comparação de itens que são invariantes em relação à localização das pessoas. Essa é uma propriedade matemática exclusiva dos modelos Rasch.

Como aponta Andrich (2004), nos modelos Rasch não há informação a mais no padrão das respostas uma vez que diferentes padrões de resposta possuem diferentes probabilidades, sendo uma fonte de verificação de desajuste ao modelo. Já nos modelos de dois e três parâmetros de TRI, diferentes padrões de resposta levam a diferentes estimativas de habilidade (Andrich, 2004). Como consequência, nos modelos Rasch, as curvas características dos itens são paralelas, o que significa que há uma invariância na ordem de dificuldade dos itens ao longo do traço latente (ou ao longo das habilidades). Portanto, itens mais fáceis para pessoas com baixa habilidade também são mais fáceis para pessoas com habilidade mediana ou alta. Já nos modelos de dois e três parâmetros, as curvas características dos itens não são paralelas, o que significa que não há invariância na ordem de dificuldade dos itens. Portanto, itens que são mais fáceis para pessoas de baixa habilidade podem se tornar mais difíceis para pessoas com maior habilidade (ver Andrich, 2004).

Uma das primeiras evidências de que o modelo Rasch é um caso especial da TMAC foi elaborada por Perline, Wright e Wainer (1979), mas a prova matemática definitiva foi apresentada recentemente por Newby, Conner, Grant e Bunderson (2009).

Em suma, alertamos que, se os avanços significativos realizados no século XX viabilizaram a produção de medidas verdadeiras em Ciências Humanas, ao mesmo tempo, tornou-se extremamente relevante que a área utilize esses avanços. Se, em alguns casos, a produção de uma medida verdadeira pode ser apenas uma opção, em avaliações *high stake* ela deveria ser indispensável. Certamente este é o caso do Exame Nacional do Ensino Médio (Enem), pois as quantificações provenientes de suas provas geram consequências sociais diretas e impactantes tanto para milhões de estudantes brasileiros quanto para as escolas de Ensino Médio, que frequentemente são avaliadas por meio dos escores de seus estudantes no Exame (Travitzki, 2013).

Atualmente, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep, 2012), autarquia do Ministério da Educação responsável pelo Enem, adota o modelo logístico de três parâmetros (3PL) da TRI para modelar a medida dos domínios latentes do Enem. Epistemologicamente, os modelos Rasch e o modelo adotado pelo Enem são muito distintos. Enquanto o modelo Rasch segue os pressupostos da TMAC e objetiva testar o quão bem os dados empíricos se ajustam aos requisitos de uma medida verdadeira, o modelo da TRI adotado pelo Enem busca criar uma modelagem capaz de explicar as propriedades presentes nos dados, adicionando ao modelo parâmetros que maximizem seu ajuste e que bem repre-

Psicologia: Teoria e Prática, *23*(1), 1-22, São Paulo, ePTPPA12625, 2021.
ISSN 1980-6909 (*on-line*)

sentem a estrutura dos dados (por exemplo, discriminação e acertos ao acaso). Bond e Fox (2015) resumem essa diferença epistemológica ao classificar o modelo de Rasch como confirmatório e preditivo, enquanto o modelo de TRI adotado pelo Enem seria um modelo exploratório e descritivo, visando o máximo de ajuste possível aos dados.

A despeito de não utilizar os modelos Rasch para analisar as quantificações produzidas, o Enem afirma que seus escores são medidas dos domínios de Linguagens, Matemática, Ciências da Natureza e Ciências Humanas. Conforme argumentamos ao longo deste texto, uma medida verdadeira é sustentada em Ciências Humanas a partir do arcabouço conceitual da TMAC e sua testagem via os modelos Rasch. Nesse sentido, até o presente momento, não sabemos se de fato o Enem mede os domínios, como se propõe, ou apenas produz meras quantificações. As implicações são extremamente relevantes. Sem uma medida verdadeira, não é possível assumir que as quantificações geradas são independentes dos itens da prova aplicada ou independentes dos indivíduos que a realizaram. Conforme explicamos, uma medida verdadeira necessita apresentar essa independência. Essa condição, ademais, é longamente reconhecida na área da psicometria e Thurstone (1931) já discutia extensivamente essa necessidade desde o início do século XX.

Considerando o exposto, este artigo tem como objetivo verificar se o Enem gera, de fato, medidas verdadeiras. Para isso, aplicamos o modelo Rasch para dados dicotômicos nos dados referentes aos acertos e erros dos estudantes nos 180 itens da edição de 2011 do Exame. Esse modelo é corretamente utilizado apenas quando os dados analisados são unidimensionais: no caso de uma prova, seus itens precisam ser majoritariamente explicados por uma habilidade. Esse parece ser o caso do Enem, já que estudos anteriores demonstraram que o fator geral de desempenho dos estudantes no Enem explica a parcela majoritária e relevante da variância dos itens da prova e apresenta maior confiabilidade (Gomes, Golino, & Peres, 2016, 2018). É importante apontar que a principal evidência dos estudos anteriores é que, ao se controlar o efeito do fator geral de desempenho no Enem (por meio de um modelo bifatorial), não apenas o ajuste aos dados é mais adequado, mas a fidedignidade composta do fator geral mantém-se elevada enquanto a dos fatores educacionais específicos ficam muito baixas (Gomes et al., 2016, 2018), o que torna a análise em separado das provas educacionais por conteúdo problemáticas. Ademais, em termos práticos, este escore geral é o escore de fato determinante para o ingresso dos estudantes nas universidades públicas brasileiras e, portanto, o que de fato provoca maior impacto social. Isso porque as universidades costumam adotar como critério para seleção dos candidatos a média dos escores nos quatro domínios avaliados (isto é, Matemática, Linguagens, Ciências da Natureza e Ciências Humanas) e na redação. Ou seja, apesar de o Inep não calcular nenhuma medida referente a um escore geral, esta é a informação que parece mais comumente ser utilizada nos sistemas de seleção ao ingresso no Ensino Superior.

2. Método

2.1 Participantes

Foram analisados os escores de 66.880 estudantes que participaram das provas do Enem de 2011 e que completaram os cadernos 120, 124, 125 e 129. Os dados foram obtidos por meio dos microdados disponibilizados publicamente pelo Inep (2012).

2.2 Instrumento

A prova de 2011 do Enem é composta por 180 itens separados em quatro grupos de 45 itens referentes aos quatro domínios (isto é, constructos ou traços latentes) avaliados pelo Exame: 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). Todos os itens são de múltipla escolha, resultando em dados dicotômicos (isto é, acerto ou erro). O banco de dados utilizados no presente estudo é o mesmo utilizado em estudos anteriores que verificaram a existência de um fator geral de desempenho (Gomes et al., 2016, 2018).

2.3 Procedimentos

O *download*, a extração, a importação e o tratamento inicial dos dados foram realizados por meio do pacote *ENEM* (Golino, 2014). Os participantes ausentes nas provas foram excluídos das análises. Em seguida, o escore dicotômico em cada item de cada prova foi calculado corrigindo-se as respostas a partir do gabarito. Os dados faltantes foram transformados em zero para as análises deste estudo.

Psicologia: Teoria e Prática, *23*(1), 1-22, São Paulo, ePTPPA12625, 2021.
ISSN 1980-6909 (*on-line*)

2.4 Análise de Dados

O modelo Rasch para dados dicotômicos foi aplicado por meio do pacote *eRm* (Mair, Hatzinger, & Maier, 2015) do R (R Core Team, 2014). De forma a verificar o ajuste dos itens ao modelo dicotômico de Rasch, utilizaram-se os índices *outfit mean square* e *infit mean square* (daqui por diante chamados apenas de *outfit* e *infit*), e o teste da razão de verossimilhança de Andersen (1973).

O *outfit* é um índice de ajuste computado a partir da média do quadrado dos resíduos padronizados de um item. Já o *infit* é um índice de ajuste que balanceia o resíduo padronizado pela variância desse resíduo e, depois, divide esse resultado pela média da variância do resíduo (Marais, 2015). Dessa forma, o *infit* não penaliza os itens que se encontram localizados longe das pessoas no contínuo da variável latente. A interpretação do (e a predileção pelo) uso do *infit mean square* é que, se um item se encontra longe da habilidade das pessoas no contínuo do traço latente, o problema não se encontra na qualidade do item na mensuração do construto, e sim na característica da amostra empregada. Dessa forma, se um item é mais difícil do que a habilidade de todas as pessoas da amostra estudada, o *outfit* irá penalizar o ajuste do item, mas o *infit* não. Nesse caso, o que o *outfit* aponta é a necessidade de encontrar pessoas com maior habilidade para aplicar o item. Da mesma forma, se o item é mais fácil do que a habilidade de todas as pessoas da amostra, o *outfit* irá penalizar o ajuste do item, indicando ser necessário encontrar pessoas com menor habilidade para aplicar o item.

Valores de *outfit* e *infit* entre 0,7 e 1,3 são considerados suficientes, mas a faixa de 0,8 a 1,2 indica um bom ajuste (Marais, 2015). Tanto o *outfit* quanto o *infit* possuem valor esperado de 1,0. Valores inferiores a 1,0 indicam que o padrão de resposta das pessoas ao item se ajusta mais do que o esperado pelo modelo. De forma semelhante, valores superiores a 1,0 indicam que o padrão de resposta das pessoas ao item se ajusta menos do que o esperado. Os índices de *infit* e *outfit* também indicam a discriminação dos itens. Itens que discriminam menos que a média de discriminação de todos os itens possuem valores de *infit* e *outfit* superiores a 1,0 (Marais, 2015). Itens que discriminam mais que a média de discriminação dos itens terão valores de *infit* e *outfit* menores que 1,0.

Já o teste da razão de verossimilhança de Andersen (1973) avalia o princípio subjacente de que, em subgrupos disjuntos arbitrários de pessoas, a estimativa do parâmetro dos itens é a mesma (hipótese nula). Dessa forma, se refutada a hipótese nula de que a estimativa do parâmetro dos itens é a mesma para *k* subgrupos, essa será uma evidência de desajuste dos itens ao modelo dicotômico de Rasch. Para computar o teste da razão de verossimilhança de Andersen, a amostra do presente estudo foi separada em quatro subamostras aleatórias.

Além do *outfit*, do *infit* e do teste de Andersen, outro indicador de qualidade é a confiabilidade de separação das pessoas e a confiabilidade de separação dos itens. Ambos são calculados por meio da relação entre a variância do erro padrão do parâmetro e o erro quadrado médio do parâmetro (MSE):

$$\textit{Confiabilidade de Separação das Pessoas} = \frac{var(\textit{Erro–Padrão de } \beta)-(\textit{MSE } \beta)}{var(\textit{Erro–Padrão de } \beta)} \tag{4}$$

$$\textit{Confiabilidade de Separação dos Itens} = \frac{var(\textit{Erro–Padrão de } \delta)-(\textit{MSE } \delta)}{var(\textit{Erro–Padrão de } \delta)} \tag{5}$$

O valor da confiabilidade de separação das pessoas e dos itens tem a mesma interpretação que o valor da confiabilidade indicada pelo alfa de Cronbach. Quanto mais próximo de 1,0, maior é a confiabilidade da medida. No entanto, esses coeficientes são interpretados no sentido de quão bem o padrão de respostas das pessoas, ou o padrão de acerto dos itens, ajusta-se à estrutura da medida. Em outras palavras, a confiabilidade de separação das pessoas indica qual a confiança que se tem de que uma pessoa que obtém uma habilidade estimada β_2 de fato possui maior habilidade do que uma outra pessoa que tenha obtido uma habilidade estimada β_2, sendo $\beta_2 > \beta_1$. De forma semelhante, a confiabilidade de separação dos itens indica qual a confiança que se tem de que um item de dificuldade estimada $\delta_2 n$ de fato possui maior dificuldade do que outro item de dificuldade estimada δ_1, sendo $\delta_2 > \delta_1$.

3. Resultados

O *infit* dos 180 itens analisados variou entre 0,81 e 1,21, apresentando média de 0,99 e desvio padrão de 0,09. Já o *outfit* variou entre 0,71 e 1,65, com média de 1,02 e desvio padrão de 0,15. No que diz respeito ao *infit* dos itens, todos os 180 itens apresentaram valores dentro da faixa de referência, entre 0,70 e 1,30 (Marais, 2015). No entanto, alguns itens apresentaram valor de *outfit* fora da faixa de referência (as letras representam os domínios teóricos e os algarismos, o número do item, com o valor do *outfit* entre parênteses): CN25 (1,66), MT20 (1,54), CN39 (1,46), CN33 (1,36), CN14 (1,34), CN8 (1,33), LC33 (1,33), CN3 (1,33), CH22 (1,33), MT33 (1,32), CN19 (1,31). Esses valores de *outfit* indicam que esses itens discriminam menos que a média de discriminação de todos os itens da prova do Enem analisada. Os padrões de resposta a esses itens são menos previsíveis do que o esperado pelo modelo Rasch. Apesar de se encontrarem fora da faixa de referência, entre 0,70 e 1,30, esses itens apresentaram valores de *infit* adequados.

O teste da razão de verossimilhança de Andersen apontou que não é possível refutar a hipótese nula de que a estimativa do parâmetro dos itens é a mesma para quatro subconjuntos aleatórios da amostra (LR = 513,022; Graus de Liberdade = 537; p = 0,76). Por sua vez, a confiabilidade de separação das pessoas foi de 0,95, enquanto a confiabilidade de separação dos itens foi de 0,99.

No que diz respeito às dificuldades dos itens, estes variaram entre -2,91 e 2,39 *logits* (M = 0; DP = 0,92). Apesar de o presente estudo empregar o modelo unidimensional de Rasch para dados dicotômicos, verificando, portanto, a variável latente do desempenho escolar geral, é interessante verificar o padrão de dificuldade dos itens do Enem por domínio escolar, uma vez que os itens são construídos seguindo uma orientação teórica que engloba quatro domínios (isto é, CN, CH, LC e MT). Os itens construídos dentro do domínio das CN apresentaram dificuldades que variaram entre -2,75 e 2,39 *logits* (M = 0,42; DP = 0,96), os de CH ficaram entre -2,91 e 1,43 *logits* (M = -0,29; DP = 0,89), os de LC variaram entre -1,91 e 1,24 *logits* (M = -0,445; DP = 0,75) e os de MT estenderam-se entre -1,32 e 1,64 *logits* (M = 0,30 DP = 0,75). A dificuldade estimada dos itens, por domínio escolar, está representada na Figura 3.1, assim como seus intervalos de confiança de 95%. A Figura 3.2, por sua vez, apresenta a distribuição das habilidades das pessoas e das dificuldades dos itens das provas do Enem de 2011.

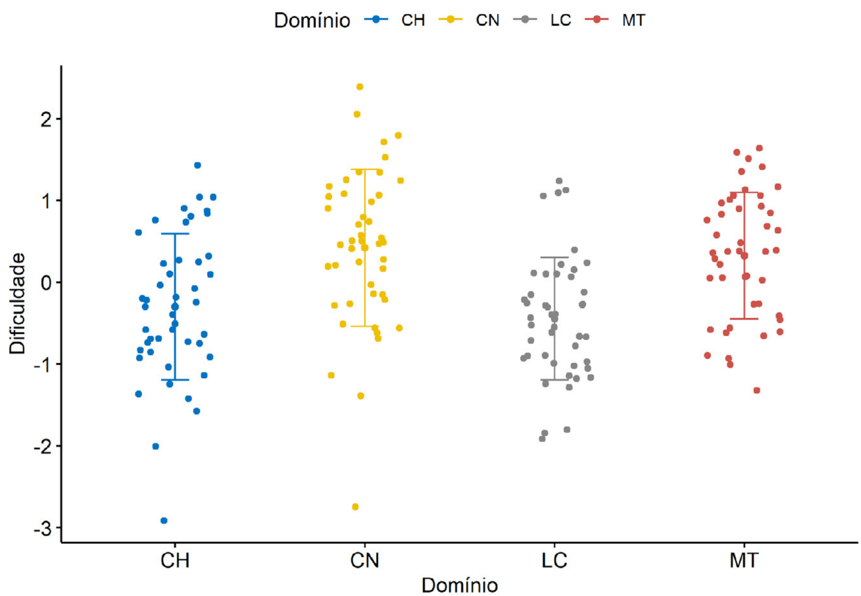

Figura 3.1. Dificuldade estimada dos itens da prova de 2011 do Enem por domínio escolar, com intervalo de confiança de 95% das dificuldades.

Legenda: CH (Ciências Humanas e suas Tecnologias); CN (Ciências da Natureza e suas Tecnologias); LC (Linguagens, Códigos e suas Tecnologias) e MT (Matemática e suas Tecnologias).

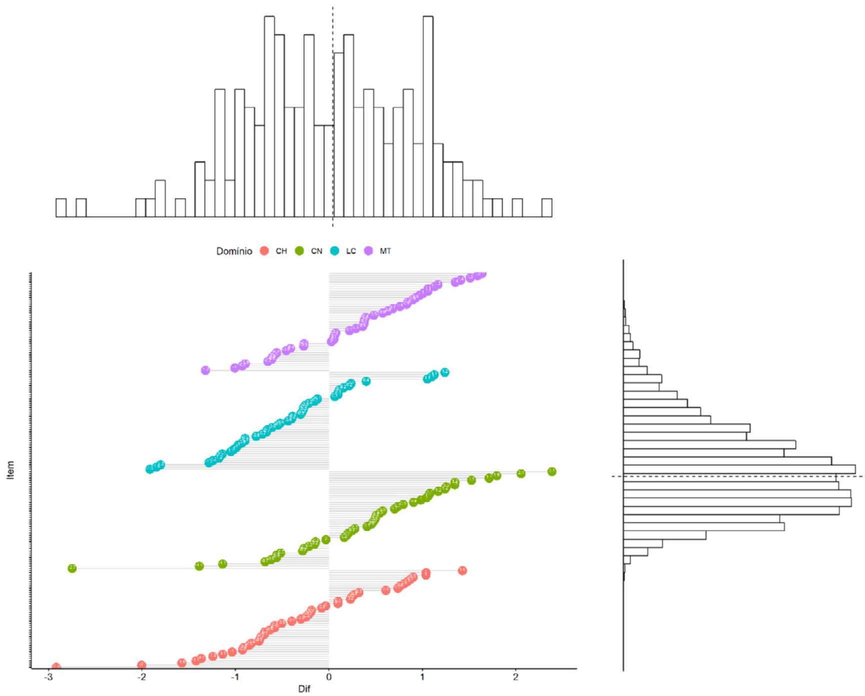

Figura 3.2. Distribuição das habilidades das pessoas (histograma do lado direito do gráfico), das dificuldades dos itens (histograma da parte superior do gráfico) e da dificuldade dos itens por domínio (centro).

Legenda: CH (Ciências Humanas e suas Tecnologias); CN (Ciências da Natureza e suas Tecnologias); LC (Linguagens, Códigos e suas Tecnologias); MT (Matemática e suas Tecnologias); Dif (dificuldade).

4. Discussão

Os resultados apontaram que os itens se ajustaram de forma adequada ao modelo Rasch, ao se considerar o índice de ajuste *infit*, com confiabilidade muito elevada de separação das pessoas e dos itens (0,95 e 0,99). Além do índice de ajuste e da confiabilidade de separação, o ajuste dos dados ao modelo Rasch foi verificado por meio do teste de razão da verossimilhança de Andersen, que revelou que os parâmetros dos itens são iguais em diferentes subgrupos de amostras. Quanto às dificuldades estimadas dos itens, verificou-se que eles compreendem quase todo o espectro de habilidades dos testandos. No entanto, há um pequeno grupo de pessoas cujas habilidades (superiores à 2,5 *logits*) não foram estimadas de forma confiável, uma vez que não há itens com dificuldades suficientemente elevadas para conseguir estimá-las. Em suma, os 180 itens do Enem de 2011 apresentam uma qualidade suficiente para se obter uma medida verdadeira do desempenho geral dos estudantes.

Apresentamos algumas implicações desses resultados. Do ponto de vista epistemológico em psicometria, é possível afirmar que a edição de 2011 do Enem atende aos axiomas da Teoria da Medida Aditiva Conjunta (TMAC), consistindo em uma medida verdadeira ou fundamental. Essa constatação confere a esse Exame maior segurança quanto ao seu modelo de medida, algo crucial considerando que o Enem é um teste *high stake*, ou seja, com impacto na vida de milhões de brasileiros e nas políticas da Educação Básica e Superior.

No entanto, é preciso relativizar esse resultado quanto ao seguinte ponto. Neste estudo, analisamos um fator geral de desempenho, de nível superior aos quatro domínios teóricos (isto é, CN, CH, MT e LC) que compõem o Exame. Estudos anteriores revelaram que o modelo bifatorial apresenta melhor ajuste aos dados do que o modelo de fatores não correlacionados atualmente adotado pelo Enem (Gomes et al., 2016, 2018). Além disso, esse fator geral é o único a apresentar confiabilidade superior a 0,95 (Gomes et al., 2016, 2018). Assim, o presente estudo corrobora que a adição de um fator geral ao modelo teórico do Enem, além de aumentar a explicação da variância dos resultados, contribui para a qualidade do instrumento de medida. Os resultados deste estudo também dão suporte à prática adotada por muitas Instituições de Ensino Superior de utilizar a média dos quatro domínios como critério para seleção de estudantes.

É importante relembrar que as características matemáticas da TMAC se aplicam unicamente aos modelos Rasch. Não obstante, o Inep, instituição responsável pelo desenvolvimento, aplicação e cálculo dos escores do Enem, utiliza o modelo de três parâmetros da Teoria de Resposta ao Item (TRI). Tanto o modelo de dois parâmetros quanto o de três não possibilitam a obtenção de uma medida verdadeira ou fundamental, pois não são aditivos (Borsboom, 2005). Ou seja, não é possível que esses modelos satisfaçam os pressupostos da TMAC. O que esses modelos fazem é modelar ou explicar o conjunto de dados (Andrich, 2004).

Como buscamos argumentar, há uma diferença grande entre modelar e medir. O primeiro tenta verificar como o dado se comporta, escolhendo-se o modelo que melhor se ajusta aos dados, no sentido que melhor o descreve. Logo, é um procedimento dado-dependente (Andrich, 2004). Já a mensuração busca identificar anomalias no dado que o fazem se distanciar de um critério operacional, matematicamente bem definido, ao qual o dado deveria se ajustar. Não havendo ajuste dos dados ao critério operacional de medida, novos dados são obtidos, e esse procedimento é repetido até que os dados se ajustem ao modelo. Como Andrich (2004, p. 13) argumenta, "identificar anomalias substantivas a partir da análise de desajuste, resistindo à modificação do modelo, [e] coletando novos dados guiados pelo modelo é consistente com o papel da medida nas ciências físicas como enunciado por Kuhn...". Leitores interessados nessa discussão podem consultar o trabalho de Andrich (2004), que elenca todos os motivos que tornam os modelos de Rasch diferentes dos modelos de dois e três parâmetros da TRI do ponto de vista da medida.

A obtenção de medidas verdadeiras na Educação e na Psicologia é relevante a partir do momento em que se pretende fazer comparações entre diferentes indivíduos, de modo a elaborar decisões relacionadas à seleção das pessoas com base no seu desempenho na avaliação. Para que esse processo seja tecnicamente justo, há que se utilizar modelos que tenham um critério matemático suportando a separação entre a habilidade das pessoas e os itens constituintes da avaliação, e os únicos modelos que têm essa propriedade são os modelos Rasch.

Em outras palavras, a comparação entre duas pessoas, em termos de suas habilidades, não deve ser afetada pelos itens que compõem o instrumento avaliativo. Essa invariância pode ser checada por meio da comparação dos parâmetros em diferentes grupos de uma amostra, como é feito usualmente nos modelos de dois e três parâmetros da TRI (Andrich, 2004). No entanto, nesses modelos da TRI, a invariância não é uma característica matemática, mas sim uma verificação empírica. Por esse motivo, essa estratégia leva a situações na análise de dados que contradizem a própria definição de invariância, uma vez que itens mais fáceis para pessoas com baixa habilidade podem ser estimados como sendo mais difíceis para pessoas com alta habilidade, colapsando o sistema de mensuração, já que a ordem da dificuldade dos itens pode se inverter em subgrupos distintos (Andrich, 2004). Essa situação gera uma incongruência do processo de mensuração injusto em um contexto de avaliações de *high stake*.

5. Conclusão

Como relatamos, os resultados apontaram que o Enem atende aos pressupostos da TMAC, quando considerado o fator geral de desempenho. Esse resultado é uma evidência favorável ao uso pelas Instituições de Ensino Superior da média dos escores nos quatro domínios específicos para a seleção de estudantes. Além disso, deve ser tratado como um indicativo da pertinência, do ponto de vista psicométrico e pedagógico, para que o Inep passe a considerar o fator geral ao divulgar os resultados do Enem.

Por fim, esperamos que este estudo sirva aos propósitos de divulgar o debate epistemológico ora apresentado a outros pesquisadores das áreas de psicometria e avaliação educacional e psicológica. Buscamos evidenciar que, ao construir instrumentos de medida em psicologia, educação e áreas afins, é necessário não apenas identificar um modelo psicométrico que melhor descreva as respostas aos itens, mas ir além da modelagem dos dados e investigar se os pressupostos epistemológicos da medida verdadeira ou fundamental estão sendo atendidos. Testes tão importantes como o Enem devem ser sistematicamente submetidos ao escrutínio de modelos que testem a qualidade das quantificações e sua viabilidade para a geração de medidas verdadeiras para que haja segurança da qualidade, do significado e da justiça das medidas por eles produzidos.

Referências

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123–140. doi:10.1007/BF02291180.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, 42(1), 7–16. doi:10.1097/01.mlr.0000103528.48582.7c.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3. ed.). London: Routledge.

Borsboom, D. (2005). *Measuring the mind*: Conceptual issues in contemporary psychometrics. New York: Cambridge University Press. doi:10.1017/CBO9780511490026.

Campbell, N. R. (1920). *Physics, the elements*. Cambridge: Cambridge University Press.

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.

Psicologia: Teoria e Prática, *23*(1), 1-22, São Paulo, ePTPPA12625, 2021.
ISSN 1980-6909 (*on-line*)

Golino, H. F., & Gomes, C. M. (2015). Teoria da medida e o modelo Rasch. In H. F. Golino, C. M. Gomes, A. Amantes, & G. Coelho. (Ed.), *Psicometria contemporânea: Compreendendo os Modelos Rasch* (pp. 13–41). São Paulo: Casa do Psicólogo/Pearson.

Gomes, C. M. A., Golino, H. F., & Peres, A. J. S. (2016). Investigando a validade estrutural das competências do ENEM: Quatro domínios correlacionados ou um modelo bifatorial. *Boletim na Medida*, *5*(10), 33–38. Recuperado de http://download.inep.gov.br/publicacoes/boletim_na_medida/2016/Boletim_Na_Medida_10.pdf

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. doi:10.5281/zenodo.2527904.

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

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement*, Vol. I. New York: Academic Press.

Mair, P., Hatzinger, R., & Maier M. J. (2015). *eRm: Extended Rasch Modeling* (Versão 0.15-5) [Software]. Recuperado de https://cran.r-project.org/web/packages/eRm/

Marais, I. (2015). Implications of removing random guessing from Rasch item estimates in vertical scaling. *Journal of Applied Measurement*, *16*(2), 113–28.

Newby, V. A., Conner, G. R., Grant, C. P., & Bunderson, C. (2009). The Rasch model and additive conjoint measurement. *Journal of Applied Measurement*, *10*(4), 348–354.

Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*(2), 237–255. doi:10.1177/014662167900300213.

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Recuperado de http://www.R-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Paedagogiske Institut.

Thurstone, L. L. (1931). The measurement of social attitudes. *The Journal of Abnormal and Social Psychology*, *26*(3), 249–269. doi:10.1037/h0070363.

Travitzki, R. (2013). *ENEM: Limites e possibilidades do Exame Nacional do Ensino Médio enquanto indicador de qualidade escolar*. (Tese de Doutorado não publicada). Faculdade de Educação, Universidade de São Paulo, São Paulo, Brasil.

Wright, B., & Stone, M. (1999). *Measurement essentials*. Wilmington: Wide Range, Inc.

Nota dos autores

Hudson F. Golino, Departamento de Psicologia, Universidade da Virgínia (UVA); **Cristiano Mauro A. Gomes,** Programa de Pós-Graduação em Neurociências (PPG Neurociências) e Programa de Pós-Graduação em Psicologia, Cognição e Comportamento (PPG Psi CogCom); Universidade Federal de Minas Gerais (UFMG); **Alexandre José de S. Peres,** Programa de Pós-Graduação em Psicologia (PPGPSICO); Universidade Federal de Mato Grosso do Sul (UFMS).

Este artigo provém de um projeto financiado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), por meio do Edital Chamada Pública Inep/Dired nº 05/2012. Cristiano Mauro Assis Gomes é bolsista de produtividade do CNPq, nível 2.

Correspondências sobre este artigo devem ser enviadas para Cristiano Mauro Assis Gomes, Universidade Federal de Minas Gerais, Departamento de Psicologia, gabinete 4036, Campus Pampulha, Av. Antônio Carlos, 6627, Pampulha, Belo Horizonte, Minas Gerais, MG, Brasil. CEP 31270-901.

E-mail: cristianomaurogomes@gmail.com