ARTIGO

Medidas de qualidade de ajuste e de DIF (*differential item functioning*)

Ruben Klein [a]

Resumo

Este artigo revê algumas medidas de qualidade de ajuste de um modelo TRI (Teoria de Resposta ao Item) e introduz as estatísticas MaxAdif e RMSD (*root mean square deviation* – raiz quadrada da média dos quadrados dos desvios) baseada nas proporções esperadas por alternativa e nas proporções empíricas por alternativa. Estende estas estatísticas para o DIF (*differential item functioning* – comportamento diferencial do item) entre dois grupos. O item pode ser dicotômico ou politômico. Indica também como calcular as proporções esperadas.

Palavras-chave: Qualidade de Ajuste. DIF. TRI. MaxAdif. RMSD. Proporções Esperadas. Proporções Empíricas.

1 Introdução

A década de 1990 representou um marco para as políticas educacionais brasileiras com a criação de diversos sistemas de avaliação da Educação Básica e da Educação Superior. Foi nessa década que foram idealizadas e executadas as principais avaliações em larga escala nacionais existentes e que geraram frutos para a criação de diversos sistemas de monitoramento e avaliação de políticas públicas em Educação nos Estados e Municípios brasileiros. Desse período destacam-se o Sistema Brasileiro de Avaliação Educacional (Saeb) e o Exame Nacional do Ensino Médio (Enem) na Educação Básica e o Exame Nacional de Cursos (ENC/Provão), que mais tarde viria a sofrer modificações e se tornar o Exame Nacional de Desempenho de Estudantes (Enade), na Educação Superior.

[a] Fundação Cesgranrio, Rio de Janeiro, RJ, Brasil.

Recebido em: 04 nov. 2024
Aceito em: 14 nov. 2024

O Saeb, em sua concepção, adotou um método inovador para a época que permitia medir, em uma mesma escala, o desempenho dos estudantes e o do item. Ademais, a adoção da Teoria da Resposta ao Item (TRI) também possibilitou a comparação do desempenho dos estudantes ao longo do tempo e entre anos escolares distintos. Posteriormente, em 2009, a TRI passou a ser utilizada também no Enem e no Exame Nacional para Certificação de Competências de Jovens e Adultos (Encceja).

A TRI pode ser compreendida como um conjunto de modelos estatísticos que modelam as respostas a um item de teste ou questionário em função de uma ou mais variáveis latentes e de características do item. Esse traço latente deve ser interpretado no contexto para o qual o instrumento foi desenvolvido. Por exemplo, em um teste cognitivo para medir o desempenho educacional e em um modelo unidimensional, pode ser interpretado como a proficiência do aluno e em contexto de questionário sobre posse de bens, pode ser interpretado como um índice socioeconômico.

A TRI pode ser utilizada tanto para itens dicotômicos quanto para itens politômicos. Nos testes educacionais, para itens dicotômicos, os modelos mais utilizados são os modelos logísticos de 1 parâmetro (Rasch), de dois parâmetros e o de 3 parâmetros (Andrade; Tavares; Valle, 2000, Baker, 1992, Birnbaum, 1968). Os principais sistemas de avaliações brasileiros até agora têm usado itens cognitivos de múltipla escolha, com apenas uma resposta correta, para acessar o desempenho dos participantes e o modelo TRI utilizado, via de regra, tem sido o de 3 parâmetros. Recentemente itens de respostas construídas têm sido acrescidos às avaliações, corrigidos com 2 ou mais categorias ordinais, e que podem ser modelados por extensões dos modelos para itens dicotômicos, como os modelos de crédito parcial, crédito parcial generalizado e o modelo graduado de Samejima (Baker, 1992). Nas avaliações nacionais, tem sido utilizado o modelo graduado de Samejima.

Em todo tipo de modelagem estatística faz-se necessário uma análise sobre os modelos para avaliar a qualidade do ajuste (em inglês, *goodness-of-*fit) obtido. De um modo geral a verificação dessa qualidade é realizada através de uma análise dos resíduos, ou seja, isto é, a diferença obtida entre o valor observado e o valor esperado. Mensurar a qualidade do ajuste do modelo é vital para uma boa análise estatística. Uma estatística de qualidade de ajuste pode possuir uma distribuição de probabilidade associada e pode ser utilizada para testar a hipótese de um bom ajuste do mesmo. Assim se não houver rejeição da hipótese do modelo estar bem ajustado então pode-se ter confiança acerca da validade das inferências que poderão ser feitas a partir do modelo ajustado.

Em particular nos modelos da TRI isso se faz ainda mais necessário quando se tem um mesmo item sendo aplicado a diversas populações, por exemplo, de anos escolares diferentes ou entre anos de aplicação distintos. Oliveri e Von Davier (2011) pontuam que diversos fatores como a diversidade social, econômica e cultural dos participantes e a familiaridade com o conteúdo dos testes pode afetar a comparabilidade desses resultados entre populações distintas.

Nos modelos da TRI uma das formas de se verificar a adequabilidade do mesmo para diversas populações é através das medidas de qualidade de ajuste para cada população e do comportamento diferencial do item (em inglês, *Differential Item Functioning – DIF)* entre as populações.

O DIF ocorre quando indivíduos de mesma proficiência em dois grupos apresentam um comportamento diferente de resposta. Pode-se investigar se um item apresenta DIF entre dois grupos comparando-se as curvas características observadas (empíricas) para os dois grupos.

Para que haja uma única escala ao longo do tempo e entre anos escolares distintos são utilizados itens para ligar os cadernos de provas. Tais itens, denominados itens âncoras, são utilizados entre os pré-testes e as provas / exames. Neste subconjunto de itens é imperativo que sejam analisadas a qualidade de ajuste e a presença do DIF seja ao longo do continuum seja entre anos escolares[1].

Estes testes estatísticos costumam ser afetados pelo tamanho da amostra. Na seção 2, introduzimos a estatística RMSD (*Root Mean Square Deviation*) para itens dicotômicos e polítomos, que não depende do tamanho da amostra. Essa estatística deve ser usada com as proporções esperadas e complementadas com as proporções empíricas.

Na seção 3 mostraremos como essa estatística complementa a metodologia que estamos usando no Saeb. As proporções esperadas de acerto costumam ser exibidas somente no BilogMG para o caso dicotômico, mas podem ser calculadas mesmo para o caso politômico. Mostraremos também como usar as proporções empíricas de acerto e por categoria de resposta.

Mesmo no caso em que os parâmetros são conhecidos e aplicados diretamente no cálculo das estimativas de proficiências de uma população, pode-se calcular

[1] Para compreender melhor como funciona o processo de calibração e equalização do Saeb recomenda-se a leitura de Klein (2003, 2009).

as proporções esperadas desde que se consiga estimar a média e desvio padrão do grupo, o que pode ser feito, por exemplo com o uso do *mirt*.

Argumentamos que essas estatísticas devem ser acompanhadas por gráficos que ajudam no entendimento da qualidade de ajuste.

Na seção 4, estenderemos o uso dessas estatísticas para o estudo de DIF entre duas populações.

2 Revisão da literatura e definição do RMSD

Para avaliar a qualidade do ajuste de modelos de TRI são utilizadas algumas medidas. Uma das mais difundidas, para um item dicotômico, é o qui-quadrado χ^2 de Bock (1972) e sua variante Q_1 de (YEN, 1981, 1984),

$$Q_{1i} = \sum_{j=1}^{10} \frac{N_j (O_{ij} - E_{ij})^2}{E_{ij} (1 - E_{ij})} \tag{1}$$

onde, são formados 10 grupos com a mesma quantidade de participantes, N_j é o número de participantes no j-ésimo grupo ($j \in \{1, 2, ..., 10\}$); O_{ij} é a proporção observada de participantes na célula j que acerta o item i; E_{ij} é a proporção predita de participantes na célula j que acerta o item i,

$$E_{ij} = \frac{1}{N_j} \sum_{k \in j}^{N_j} \hat{P}_i (\hat{\theta}_k) \tag{2}$$

e, $\hat{P}_i (\hat{\theta}_k)$ é o valor da curva característica do item i avaliada usando a proficiência ($\hat{\theta}_k$) estimada para o participante k e os parâmetros estimados para o item i. A quantidade Q_1 é igual ao χ^2 de Bock sob algumas condições: quando os participantes são agrupados em j partes, mas não necessariamente com $j_{max} = 10$ e além disso $E_{ij} = (\hat{P}_i) (\hat{\theta}_k)$ onde $\hat{\theta}_k$ é a mediana dos valores da habilidade estimados para os participantes no grupo j. O número de graus de liberdade associados com a estatística qui-quadrado é da ordem de $j_{max} - m$, onde m é o número de parâmetros do item. Ambas medidas estão implementadas no pacote *mirt* (Chalmers, 2012) do *software* estatístico R (R Core Team, 2020). Também implementado nesse mesmo pacote há o teste G^2 de razão de verossimilhança de McKinley e Mills (1985), similar ao Q_1.

$$G_i^2 = 2 \sum_{j=1}^{20} O_{ij} \ln\left(\frac{O_{ij}}{E_{ij}}\right) \sim \chi^2_{(10-m)} \tag{3}$$

Esses métodos podem também ser aplicados no caso politómo.

No modelo Rasch, para cada item, utilizam-se as estatísticas OUTFIT e INFIT (Wright; Masters, 1982) baseadas nos resíduos estandardizados.

De um modo geral, tais estatísticas são afetadas pelo tamanho da amostra e tendem a rejeitar a hipótese do modelo a medida que o tamanho da amostra aumenta. Destarte, foi criado a estatística *Root Mean Square Error of Approximation – RMSEA* (Steiger, 2016; Steiger; Lind, 1980) e redefinida em Tennant e Pallant (2012) como:

$$RMSD = \sqrt{\max\left[\frac{\frac{\chi^2}{df} - 1}{N - 1}; 0\right]} \tag{4}$$

onde χ^2 é a estatística qui-quadrado, *df* seu grau de liberdade e *N* o tamanho da amostra. RMSEA tem um valor esperado de 0 (zero) quando os dados se ajustam ao modelo. RMSEA também é definido em Maydeu-Olivares (2013) por:

$$RMSEA = \sqrt{\max\left[\frac{\frac{\chi^2}{df} - 1}{N}; 0\right]} \tag{5}$$

que é, praticamente o mesmo valor para N grande. Maydeu-Olivares (2013) sugere um ponto de corte para um ajuste excelente 0,05/(K-1) onde K é o número de categorias do item. Oliveri e Von Davier (2011) propõem a utilização do RMSEA no contexto de identificação da qualidade de ajuste para cada um dos 30 países membros da OCDE como uma subpopulação distinta, no Programa Internacional de Avaliação de Estudantes (Pisa) 2006. Nesse trabalho, valores de RMSEA maiores que 0,10 indicaram mau ajuste.

George et al. (2016), no contexto dos Modelos de Diagnóstico Cognitivo (em inglês, *Cognitive Diagnosis Models* – CDM), implementam uma versão discretizada dessa estatística:

$$RMSEA_j = \sum_{k=1}^{K}\sum_{d=1}^{D} \pi(a_{kd}) \left[P(X=1 \mid a_{kd,j}) - \frac{N(X=1 \mid a_{kd,j})}{N(a_{kd,j})}\right] \quad (7)$$

onde, a_{kd} é o nível de desempenho d para a dimensão k; $\pi(a_{kd})$ é a proporção de participantes pertencentes ao nível a_{kd} ; $P\ (X = 1 \mid a_{kd,j})$ é a probabilidade estimada pelo modelo para um participante pertencente ao nível a_{kd} resolver o item j; $N\ (X = 1 \mid a_{kd,j})$ é o número observado de participantes no nível a_{kd} que responderam corretamente o item j e $N(a_{kd,j})$ é o número total de respondentes ao item j no nível a_{kd}.

No estudo de George *et al*. (2016) são propostos alguns pontos de corte para essa medida. Valores de RMSEA menores do que 0,05 indicam um ajuste excelente, já valores entre 0,05 e 0,10 indicam um ajuste moderado e valores superiores a 0,10 indicaram um mau ajuste.

No R, os pacotes CDM (George *et al*., 2016; Robitzsch *et al*., 2020) e TAM (Robitzsch; Kiefer; Wu, 2020) implementam essa estatística.

Uma outra importante estatística para a qualidade do ajuste e que não depende do tamanho da amostra é o *RMSD,* utilizada no PISA 2015 (OCDE, 2017)*,* no ciclo 2012-2016 do PIAAC (Yamamoto, Khorramoel, Von Davier, 2013) and OCDE, PIAAC (OCDE, 2019) e no Pisa para Escolas (Okubo *et al.*, 2021). Esta estatística é definida por:

$$RMSD = \sqrt{\int (P_o\ (\theta) - P_e\ (\theta)^2\ f(\theta) d(\theta)} \quad (6)$$

onde $P_o\ (\theta)$ é curva característica observada do item (proporções esperadas obtidas no passo E do processo EM de estimação) e $P_e\ (\theta)$ é a curva característica do item e f(θ) é a função densidade da distribuição da proficiência θ, obtida no final do processo. Isso vale para itens dicotômicos ou para cada categoria de um item politômico ou ainda para as proporções acima de uma categoria em um item politômico ordenado com a densidade da distribuição cumulativa.

No Pisa 2015 (OCDE, 2017) são propostos vários pontos de corte para o RMSD em diferentes situações, como por exemplo para a comparação dos parâmetros entre um país e a escala global mas, para o estudo da qualidade de ajuste usa o ponto de corte de 0,12. Okubo *et al.* (2021), no Pisa para Escolas também utiliza o ponto de corte 0.12. Achamos que poderiam ser mais severos e usar 0,10, já que todos usam as proporções esperadas.

Para um item politômico com categorias 0,1,..,K, Khorramdel, Shin, and Von Davier (2019), Pisa Technical Report 2018, chapter 16) (OCDE, 2022), propõe um RMSD único da seguinte maneira:

$$RMSD = \sqrt{\left(\sum_{k=0}^{K} \int (P_{ok}(\theta) - P_{ek}(\theta)^2 f(\theta) d(\theta)\right) / (K+1)}$$

Observa-se que se $RMSD_k$ é o RMSD da categoria k e que

$$1/(K+1) \sum_{k=0}^{K} RMSD_k \leq \sqrt{(1/(K+1)) \sum_{k=0}^{K} (RMSD_k)^2}$$

Esse RMSD único pode ser estendido para o caso das K curvas acumuladas P(X≥k). k=1,..,K Nesse caso, a divisão é por K.

Observa-se que no caso de estimação por grupos múltiplos, a densidade depende do grupo.

3 Qualidade de ajuste

A qualidade de ajuste de um item dicotômico é verificada comparando-se as proporções esperadas de acerto obtidas pelo método EM de estimação em alguns pontos por exemplo, fornecidas pelo BilogMG (Du Toit, 2003) nos arquivos de saída .exp, ou as proporções empíricas de acerto com as respectivas probabilidades de acerto (os valores da curva característica do item).

O procedimento adotado nos Saebs e Enems (Relatório Enem 2016 para ajuste), propostos pelo autor, denominado MaxAdif, tem sido o de pegar o máximo dessas diferenças em um intervalo com valores próximos dos percentis 5% e 95% e usar como ponto de corte o valor 0,15. Rejeita-se a qualidade de ajuste se esse valor absoluto do máximo for maior que o ponto de corte. Esse ponto de corte pode

ser relaxado, em particular, se o máximo ocorre nos extremos. Esse valor 0,15 é utilizado tanto para as proporções empíricas de acerto quanto para as proporções esperadas fornecidas pelo método de estimação EM. No entanto a variabilidade das proporções empíricas é maior.

No caso da proporção esperada estimada pelo método EM, utiliza-se os pontos de quadratura, entre os quantis 5% e 95% da distribuição de proficiência da população e no caso das proporções empíricas utiliza-se os pontos (níveis) arbitrados, em geral, a média e os pontos espaçados por meio desvio padrão, também entre os quantis 5% e 95% da distribuição de proficiência da população.

No caso dicotômico, discretizando a equação (6), seja θ_k um ponto de quadratura, $P_j(\theta_k)$ a proporção esperada de acerto para o item j no ponto de quadratura θ_k, $P(X = 1 \mid \theta_k)$ a probabilidade dos participantes com proficiência θ_k acertarem ao item j e ω_k o peso dado por $f(\theta_k)\Delta\theta$, onde f é a densidade final calculada no ponto θ_k e $\Delta\theta$ é o intervalo entre os pontos de quadratura. Com estas notações, a eq. 6 é equivalente à:

$$RMSD = \sqrt[2]{\sum_{k=1}^{K} \omega_k \left(P_j(\theta_k) - P_j(X = 1 \mid \theta_k)\right)^2} \tag{9}$$

No caso politômico com categorias ordenadas, g = 0,1,...,G, para cada categoria g, temos $P_{jg}(\theta_k)$, no lugar de $P_j(\theta_k)$, é a proporção esperada de respostas na alternativa g e $P_j(X = g \mid \theta_k)$ é o valor da curva característica da probabilidade da categoria g. Pode-se também adaptar para a proporção empírica e a probabilidade da alternativa ser maior ou igual do que g ($\geq g$).

No caso do BilogMG (Du Toit, 2003), os arquivos .exp fornecem os pontos de quadratura (POINT), os pesos (WEIGHT), as proporções esperadas (PROPORTION) e os valores da curva característica (MODEL PROP). No caso de mais de um grupo, no BilogMG, os pontos de quadratura são os mesmos para todos os grupos, mas os pesos dependem do grupo.

Mas pode-se calcular as proporções esperadas em qualquer caso, não se dependendo do BilogMG.

A proporção esperada de respostas na categoria g do item j no ponto de quadratura θ_k é dada por r_{jgk}/N_{jk}, onde

$$r_{jgk} = \sum_{i=1}^{N} x_{igk}\, P(\theta_k \mid x_i, \xi)$$

é o número esperado de respostas na categoria g= *0*, *1*,,,.*G* do item *j* por indivíduos com proficiência no intervalo ($\Theta_k - \Delta\Theta/2$, $\Theta_k + \Delta\Theta/2$) e

$$N_{jk} = \sum_{g=0}^{G} r_{jgk}$$

é o número esperado de indivíduos que responderam ao item *j* com proficiência no intervalo ($\Theta_k - \Delta\Theta/2$, $\Theta_k + \Delta\Theta/2$).

Observa-se também que

$$P(\theta_k \mid x_i, \xi) = \left(P(x_i \mid \theta_k, \xi)\, \omega_k\right) \Big/ \sum_{k=1}^{K} P(x_i \mid \theta_k, \xi)\, \omega_k$$

é a probabilidade *a posteriori* da proficiência estar no intervalo ($\Theta_k - \Delta\Theta/2$, $\Theta_k + \Delta\Theta/2$) dado o vetor de resposta x_i e o vetor de parâmetros ξ e onde o peso ω_k é obtido a partir da distribuição final do grupo dada pelos softwares no processo de calibração.

Para o caso empírico basta assumir umas pequenas modificações à essa função de tal forma que o peso ω_{jk} seja proporcional ao número de alunos no ponto (nível) θ_k. Observa-se que o peso novamente depende do grupo e pode depender do item, no caso de haver diferentes cadernos. No segundo caso pode-se fazer o peso proporcional ao número de alunos, por ponto (nível), que respondeu ao item. $P_j\,(\theta_k)$, neste caso, seria a proporção empírica de acerto dos participantes que estão no ponto (nível) θ_k e $P_j\,(X = 1 \mid \theta_k)$ o valor da curva característica do item no ponto (nível) θ_k (probabilidade de acerto no nível θ_k).

No caso politômico com categorias ordenadas, $P_j\,(\theta_k)$ é a proporção empírica de respostas na alternativa g e $P_j\,(X = g \mid \theta_k)$ é o valor da curva característica da probabilidade da categoria g. Pode-se também adaptar para a proporção empírica e a probabilidade da alternativa ser maior ou igual do que g ($\geq g$).

Pode-se calcular também o RMSD único para o item politômico e, também, adaptar para as funções cumulativas de cada categoria partindo do 1º parcial e dividindo por G.

Continuamos chamando essa estatística de RMSD, no entanto, diferentemente da MaxADif não há necessidade de se restringir a estatística ao intervalo compreendido entre os quantis 5% e 95% uma vez que, os pesos, fora desse intervalo, tendem a ser pequenos e muitas vezes nulos no caso do cálculo baseado nos percentuais empíricos.

O *Pisa Technical report* 2015 menciona ainda a estatística MD (*mean deviation*) e nós acrescentamos a estatística MAD (*mean absolute deviation*), que podem ser definidos assim:

MD = média ponderada das diferenças entre as proporções esperadas ou empíricas e a probabilidade do modelo.

MAD = média ponderada dos valores absolutos das diferenças entre as proporções esperadas ou empíricas e a probabilidade do modelo. Observa-se que é uma versão robusta do RMSD.

O pacote R CDM também calcula as estatísticas MD e MAD.

Observamos que RMSD é a raiz quadrada da média ponderada dos quadrados das diferenças entre as proporções esperadas ou empíricas e a probabilidade do modelo.

O uso de RMSD > 0.12 ou 0.10 elimina muito mau ajuste quando MaxAdif é grande somente nos extremos ou apenas em algum ponto perto do extremo, casos que sempre precisamos decidir o que fazer.

Por outro lado, se MaxAdif > 0.10 ou 0.12 e < 0.15, mas a grande maioria das diferenças no centro for maior que 0.10 ou 0.12, RMSD pode indicar mau ajuste. É o caso de um deslocamento para a direita ou para a esquerda.

Em todos os exemplos, nos gráficos as proporções esperadas são plotadas por símbolos abertos (ou vazados) e as proporções empíricas por símbolos cheios.

As proporções empíricas foram sempre calculadas com proficiências obtidas pelo método EAP e utilizando a *priori* N(0,1). Provavelmente as proporções empíricas obtidas a partir das proficiências calculadas pelo método EAP e *a priori* como a distribuição normal com a média e desvio padrão do grupo estariam mais próximas das curvas.

Exemplo 1. Item dicotômico, em uma avaliação com 3 grupos e um grande número de indivíduos em cada grupo.

A análise foi feita com o pacote R *mirt* utilizando a rotina multipleGroup e a proficiência foi estimada pelo método EAP. O grupo 1 é o grupo do "meio" entre os grupos. O Quadro 1 apresenta quantis das proficiências dos 3 grupos, o Quadro 2 apresenta as diferenças entre as proporções esperadas e as probabilidades estimadas pelo modelo, entre os pontos próximos aos quantis 5 e 95 e as estatísticas de qualidade de ajuste MaxAdif e RMSD. O Quadro 3 apresenta as diferenças com as proporções empíricas em vez das esperadas. O Gráfico 1 mostra a curva característica do item e as proporções esperadas e empíricas nos pontos. Como se pode ver, o gráfico ajuda a interpretar as estatísticas do Quadro 2. O grupo 1 é bem ajustado, MaxAdif 0.018 e RMSD 0.011 para as proporções esperadas e MaxAdif 0.044 e RMSD 0.022 para as proporções empíricas. Os outros dois grupos não são tão bem ajustados, mas estão dentro dos limites de aceitação. No grupo 3, o RMSD do ajuste com as proporções esperadas deu 0.092, abaixo do ponto de corte de 0.10. Para o ajuste com as proporções empíricas deu 0.101 e não seria rejeitado para o ponto de corte de 0.12. O MaxAdif desse grupo foi de 0.111 para as proporções esperadas e de 0.115 para as proporções empíricas, bem abaixo de 0.15. O Gráfico 1 mostra que as "curvas características observadas esperadas e empíricas" para os grupos 2 e 3 são afastadas e quase paralelas a curva característica do item.

Quadro 1 - Quantis das distribuições de proficiências dos 3 grupos

	0%	1%	5%	10%	25%	50%	75%	90%	95%	99%	100%
grupo 2	-2.60	-2.00	-1.47	-1.16	-0.63	-0.01	0.61	1.16	1.50	2.16	2.88
grupo 1	-2.60	-2.00	-1.48	-1.17	-0.61	0.02	0.64	1.18	1.51	2.13	2.88
grupo 3	-2.60	-2.20	-1.71	-1.41	-0.82	-0.13	0.52	1.08	1.41	2.05	2.88

Fonte: Elaboração própria (2023)

Quadro 2 - Estatísticas de qualidade de ajuste pelas proporções esperadas para os 3 grupos

	-1.538	-1.128	-0.718	-0.308	0.103	0.513	0.923	1.333	MaxAdif	RMSD
grupo 2	0.023	0.046	0.066	0.078	0.080	0.075	0.066	0.056	0.080	0.066
grupo 1	0.011	0.017	0.018	0.013	0.004	-0.004	-0.009	-0.009	0.018	0.011
grupo 3	-0.065	-0.076	-0.088	-0.100	-0.109	-0.111	-0.106	-0.093	0.111	0.092

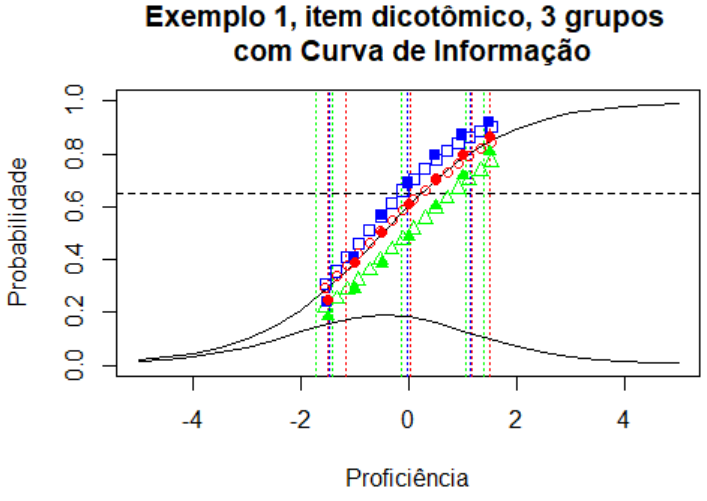
Fonte: Elaboração própria (2023)

Quadro 3 - Estatísticas de qualidade de ajuste pelas proporções empíricas para os 3 grupos

	-1.5	-1	-0.5	0	0.5	1	1.5	MaxAdif	RMSD
grupo 2	-0.050	0.024	0.072	0.088	0.093	0.088	0.073	0.093	0.078
grupo 1	-0.044	0.000	0.010	0.007	0.006	0.016	0.019	0.044	0.022
grupo 3	-0.107	-0.100	-0.110	-0.115	-0.099	-0.067	-0.041	0.115	0.101

Fonte: Elaboração própria (2023)

Gráfico 1 - Curva característica do item, com as proporções esperadas (símbolos abertos) e empíricas (símbolos cheios), os quantis 5, 10, 50, 90 e 95 dos 3 grupos, a linha horizontal no ponto 0.65 e a curva de informação



Fonte: Elaboração própria (2023)

O Quadro 4 apresenta a saída da função itemfit do *mirt* com a estatística chi-quadrada X^2, seus graus de liberdade, a estatística RMSEA e a probabilidade de significância para cada grupo. Como esperado a probabilidade de significância é zero e a hipótese do modelo ser bem ajustada aos dados rejeitada. Mas o RMSEA nos mostra outro quadro, com o grupo 1 sendo bem ajustado, e os grupos 2 e 3 com ajuste moderado, sendo o pior ajuste o do grupo3.

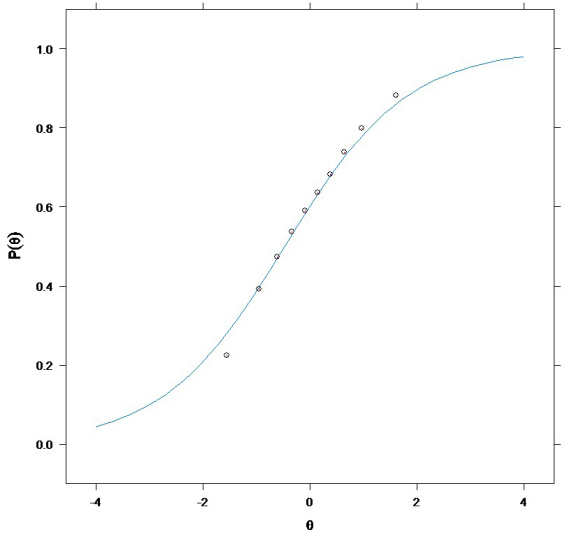
Os Gráficos 2a, 2b e 2c, mostram a curva característica do item e as proporções esperadas utilizadas no cálculo de X2. Novamente o gráfico ajuda a interpretar a estatística e se pode ver no Gráfico 2 o mesmo padrão do Gráfico 1, com o grupo 1 bem ajustado e o pior sendo o grupo 3.

Quadro 4 - Quadro de saída do *mirt* usando a função itemfit e a opção X2

	X^2	df.X^2	RMSEA.X^2	p.X^2
grupo 2	57201.162	8	0.060	0
grupo 1	4240.389	8	0.017	0
grupo 3	59555.517	8	0.072	0

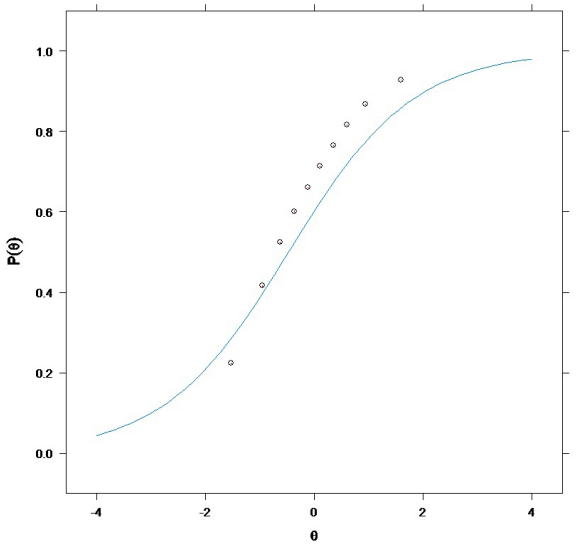
Fonte: Elaboração própria (2023)

Gráfico 2a - Curva e proporções empíricas de acerto nos pontos usados para o teste X^2, chi-quadrado, usando a função itemfit com opção empirical.plot para o grupo1



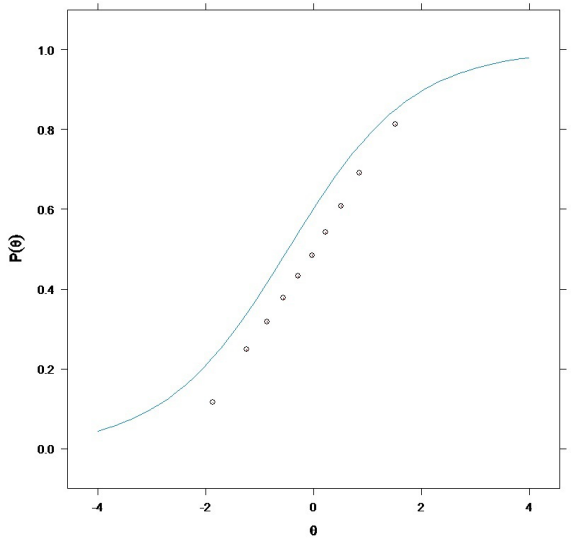
Fonte: Elaboração própria (2023)

Gráfico 2b - Curva e proporções empíricas de acerto nos pontos usados para o teste X^2, chi-quadrado, usando a função itemfit com opção empirical.plot para o grupo2



Fonte: Elaboração própria (2023)

Gráfico 2c - Curva e proporções empíricas de acerto nos pontos usados para o teste X^2, chi-quadrado, usando a função itemfit com opção empirical.plot para o grupo3



Fonte: Elaboração própria (2023)

Exemplo 2. Item politômico com 3 categorias de respostas ordenadas em uma avaliação com cerca de 13.000 indivíduos.

A análise foi feita com o pacote R *mirt* com a rotina *mirt*, com o modelo graduado de Samejima (graded) e a proficiência estimada pelo método EAP. O Quadro 5 apresenta os quantis da distribuição de proficiência, o Quadro 6 apresenta as diferenças entre as proporções esperadas e as probabilidades estimadas pelo modelo para as 3 categorias de resposta, nos pontos próximos aos quantis 5 e 95 e as estatísticas de qualidade de ajuste MaxAdif e RMSD. e o Quadro 7 apresenta as diferenças entre as proporções esperadas e as probabilidades estimadas pelo modelo para as categorias acumuladas (≥ 1 e ≥ 2) de resposta, entre os pontos próximos aos quantis 5 e 95 e as estatísticas de qualidade de ajuste MaxAdif e RMSD. Os Quadros 8 e 9 apresentam essas diferenças e estatísticas com as proporções empíricas em vez das proporções esperadas. Os RMSDs únicos são pequenos em todos os casos.

O Gráfico 3 mostra as curvas características das 3 categorias de resposta do item e as proporções esperadas e empíricas nos pontos. Como se pode ver, o gráfico ajuda a interpretar as estatísticas dos Quadros 6 e 8. A categoria 2 é bem ajustada, MaxAdif 0.041 e RMSD 0.018 no caso das proporções empíricas. São menores no caso das proporções esperadas. As outras categorias são bem ajustadas pelas proporções esperadas, mas não tão bem quando se analisa pelas proporções empíricas. Mas estão dentro dos limites de aceitação de RMSD. Os MaxAdifs empíricos dessas categorias são maiores que 0.15, mas isso ocorre perto do quantil 5%. No entanto os MaxAdifs esperados são pequenos. A análise é semelhante para os Quadros 7 e 9 e o Gráfico 4.

Quadro 5 - Quantis da distribuição de proficiências

0%	5%	10%	25%	50%	75%	90%	95%	100%
-2.660	-1.770	-1.308	-0.543	0.099	0.649	1.096	1.351	2.096

Fonte: Elaboração própria (2023)

Quadro 6 - Estatísticas de qualidade de ajuste para as proporções esperadas para as curvas características das probabilidades de resposta

	-1.744	-1.333	-0.923	-0.513	-0.103	0.308	0.718	1.128	MaxAdif	RMSD
cat 0	0.029	0.027	-0.004	-0.023	-0.018	-0.010	-0.004	-0.002	0.032	0.015
cat 1	-0.027	-0.025	-0.001	0.006	-0.006	-0.011	-0.007	-0.002	0.029	0.010
cat 2	-0.002	-0.002	0.004	0.016	0.024	0.021	0.011	0.004	0.024	0.015

O RMSD único é: 0.014
Fonte: Elaboração própria (2023)

Quadro 7 - Estatísticas de qualidade de ajuste para as proporções esperadas para as curvas características das probabilidades acumuladas de resposta (maior ou igual as categorias 1 e 2)

	-1.744	**-1.333**	**-0.923**	**-0.513**	**-0.103**	**0.308**	**0.718**	**1.128**	**MaxAdif**	**RMSD**
cat 1	-0.029	-0.027	0.004	0.023	0.018	0.010	0.004	0.002	0.032	0.015
cat 2	-0.002	-0.002	0.004	0.016	0.024	0.021	0.011	0.004	0.024	0.015

O RMSD único é: 0.015
Fonte: Elaboração própria (2023)

Quadro 8 - Estatísticas de qualidade de ajuste para as proporções empíricas para as curvas características das probabilidades de resposta

	-2	**-1.5**	**-1**	**-0.5**	**0**	**0.5**	**1**	**1.5**	**MaxAdif**	**RMSD**
cat0	0.090	0.185	0.168	-0.074	-0.031	-0.019	-0.006	-0.003	0.185	0.076
cat1	-0.081	-0.158	-0.127	0.077	0.016	0.024	0.010	-0.032	0.158	0.064
cat2	-0.008	-0.026	-0.041	-0.003	0.015	-0.005	-0.004	0.034	0.041	0.018

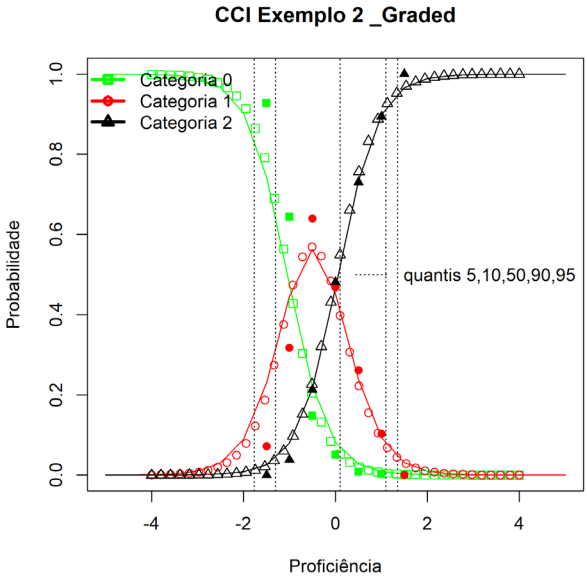
O RMSD único é: 0.058
Fonte: Elaboração própria (2023)

Quadro 9 - Estatísticas de qualidade de ajuste para as proporções empíricas para as curvas características das probabilidades acumuladas de resposta (maior ou igual as categorias 1 e 2)

	-2	**-1.5**	**-1**	**-0.5**	**0**	**0.5**	**1**	**1.5**	**MaxAdif**	**RMSD**
cat 1	-0.090	-0.185	-0.168	0.074	0.031	0.019	0.006	0.003	0.185	0.076
cat 2	-0.008	-0.026	-0.041	-0.003	0.015	-0.005	-0.004	0.034	0.041	0.018

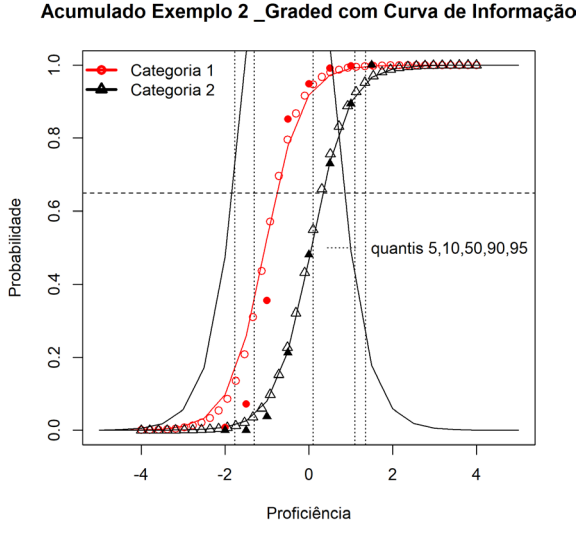
O RMSD único é: 0.055
Fonte: Elaboração própria (2023)

Gráfico 3 - Curvas características das probabilidades de resposta às categorias e as proporções esperadas e empíricas de resposta



Fonte: Elaboração própria (2023)

Gráfico 4 - Curvas características das probabilidades acumuladas de resposta às categorias e as proporções esperadas e empíricas de resposta



Fonte: Elaboração própria (2023)

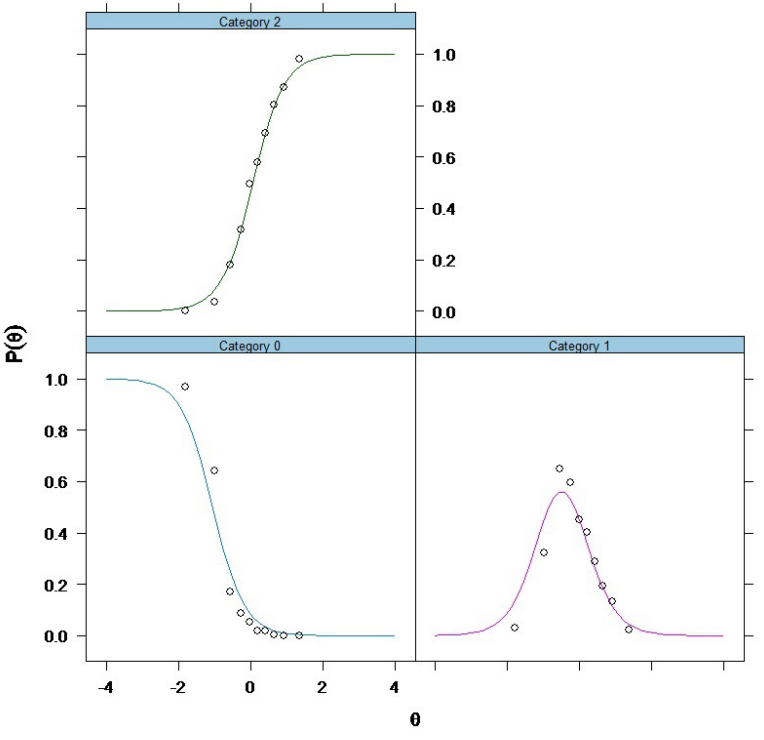
O Quadro 10 mostra a saída da função itemfit do *mirt*. O RMSEA é único, pois é baseado na estatística chi-quadrado X2. O RMSEA de 0.02 indica um bom ajuste, mas o Gráfico 5 mostra um quadro semelhante ao Gráfico 4, indicando o bom ajuste da categoria 2, e os desvios nas outras duas categorias.

Quadro 10 - Quadro de saída do *mirt* usando a função itemfit e a opção X^2

	X^2	df.X^2	RMSEA.X^2	p.X^2
74	102.227	16	0.020	0

Fonte: Elaboração própria (2023)

Gráfico 5 - Gráfico *mirt* das curvas características das probabilidades das categorias de resposta e as proporções empíricas utilizadas no cálculo de X^2



Fonte: Elaboração própria (2023)

4 DIF (*differential item functioning*)

Um item apresenta DIF (comportamento diferencial do item) se indivíduos de mesma proficiência em dois grupos distintos apresentam um comportamento diferente de resposta.

Então uma maneira de se verificar se há DIF é comparar as proporções esperadas (empíricas) nos diversos pontos. É como substituir na análise da qualidade de ajuste a curva característica do item pelas proporções esperadas (empíricas) do segundo grupo.

Como no ajuste de qualidade, o MaxAdif é definido de forma semelhante ou seja:

MaxAdif = máximo das diferenças entre as proporções esperadas (empíricas) entre o máximo dos quantis 5% dos dois grupos e o mínimo dos quantis 95% dos dois grupos.

O ponto de corte utilizado continua sendo 0.15.

A definição do RMSD é modificada para:

$$RMSD = \sqrt{\sum peso*(proporção1 - proporção2)^2}$$

É preciso especificar o peso a ser utilizado. Esse peso pode ser definido a partir dos pesos dos dois grupos, utilizados no estudo da qualidade de ajuste.

Há 3 propostas naturais. O peso é proporcional a:

a) média aritmética dos pesos de cada grupo.

b) média geométrica dos pesos de cada grupo.

c) produto dos pesos dos dois grupos.

As duas últimas são preferíveis pois se o peso em um dos grupos for zero em um ponto, o produto é zero. A terceira proposta, o produto dos pesos, tende a dar menos peso nos extremos, como fora do intervalo definido em MaxAdif,

No entanto, a média geométrica tem a propriedade adicional de que se peso1 = peso2, a média geométrica retorna este peso.

O ponto de corte para RMSD a ser utilizado continua sendo 0.12 ou 0.10.

No exemplo 1, o Quadro 11 mostra as diferenças entre as proporções esperadas de acerto do item dicotômico em alguns pontos de quadratura no intervalo definido para MaxAdif e as estatísticas RMSDP (definição c) e RMSDS (definição b) para as 3 combinações dos grupos e o Quadro 12 faz o mesmo para as proporções empíricas. Pode-se ver, pelos dois quadros que há DIF entre o grupo 2 e 3 por ambos os critérios, e que há DIF entre os grupos 1 e 3 pelo RMSD se o ponto de corte do RMSD for 0.10. O Gráfico 1 confirma essas conclusões.

Quadro 11 - Estatísticas de DIF entre os 3 grupos com as proporções esperadas

	-1.538	**-0.923**	**-0.308**	**0.308**	**0.923**	**MaxAdif**	**RMSDP**	**RMSDS**
grupos 1X2	-0.011	-0.039	-0.065	-0.079	-0.075	0.079	0.067	0.063
grupos 1X3	0.076	0.100	0.113	0.111	0.097	0.114	0.105	0.099
grupos 2X3	-0.087	-0.139	-0.178	-0.189	-0.172	0.189	0.169	0.158

Fonte: Elaboração própria (2023)

Quadro 12 - Estatísticas de DIF entre os 3 grupos com as proporções empíricas

	-1.500	**-1**	**-0.500**	**0**	**0.5**	**1**	**1.5**	**MaxAdif**	**RMSDP**	**RMSDS**
grupos 1X2	-0.006	0.023	0.062	0.081	0.087	0.072	0.053	0.087	0.072	0.067
grupos 1X3	0.063	0.101	0.119	0.122	0.105	0.083	0.060	0.122	0.109	0.102
grupos 2X3	0.057	0.124	0.181	0.203	0.192	0.155	0.113	0.203	0.178	0.165

Fonte: Elaboração própria (2023)

5 Conclusão

A introdução do RMSD fornece um outro instrumento muito útil para se estudar a qualidade de ajuste da curva característica de um item aos dados bem como

para se estudar se há DIF entre grupos. O RMSD vem sendo utilizado no PIIAC, no Pisa e no Pisa para Escolas. Nessas aplicações, para o estudo de qualidade de ajuste, tem sido usado as proporções esperadas. Nesse artigo, mostra-se como essas podem ser calculadas. Mostra-se também que é útil calcular o RMSD com as proporções empíricas e que se deve fazer os gráficos com a curva característica e as duas proporções.

Quality measures of fit and DIF (differential item functioning)

Abstract

This paper reviews some measures of quality of fit in IRT (Item Response Theory) model and introduces the statistics MaxAdif and RMSD (root mean squre deviation) based on expected proportions by alternative and on empirical proportions by alternative. The paper extends these statistics for the DIF (differential item functioning) between two groups. The item can be dichotomous or polytomous. It also indicates how to compute the espected proportions.

Keywords: *Quality of Fit. DIF. IRT. MaxAdif. RMSD. Expected Proportions, Empirical Proportions.*

Medidas de calidad de ajuste y DIF (funcionamiento diferencial del ítem)

Resumen

Este artículo revisa algunas medidas de calidad de ajuste del modelo IRT (Teoría de respuesta al ítem) y introduce las estadísticas MaxAdif y RMSD (desviación cuadrática media) basadas en proporciones esperadas por alternativa y en proporciones empíricas por alternativa. El artículo amplía estas estadísticas para el DIF (funcionamiento diferencial del ítem) entre dos grupos. El ítem puede ser dicotómico o politómico. También indica cómo calcular las proporciones esperadas.

Palabras clave: *Calidad de Ajuste. DIF. IRT. MaxAdif. RMSD. Proporciones Esperadas. Proporciones Empíricas.*

Referências

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C. Teoria da Resposta ao Item: conceitos e aplicações. *In*: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 14., 2000, Caxambu. *Livro de resumos* [...]. Caxambu: Associação Brasileira de Estatística, 2000.

BAKER, F. B. Item Response Theory: parameter estimation techniques. New York: Marcel Dekker, 1992. (Statistics, textbooks, and monographs, v. 129).

BIRNBAUM, A. Some latent trait models and their use in inferring an examinee's ability. *In*: LORD, F. M.; NOVICK, M. R. (eds.). Statistical theories of mental test scores. [S. l.] : Addison-Wesley, Reading, 1968. p. 397-479. (The Addison-Wesley series in behavioral sciences: quantitative methods)

BOCK, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, [s. l.], v. 37, n. 1, p. 29-51, mar. 1972. https://doi.org/10.1007/BF02291411

CHALMERS, R. P. mirt : A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, [s. l.], v. 48, n. 6, p. 1-29, 2012. https://doi.org/10.18637/jss.v048.i06

DU TOIT, M. *IRT from SSI*: BILOGMG, MULTILOG, PARSCALE, TESTFACT. Chicago: Scientific Software International, 2003.

GEORGE, A. C. *et al*. The R package CDM for cognitive diagnosis models. *Journal of Statistical Software*, [s. l., v. 74, n. 2, p. 1-24, 2016. https://doi.org/10.18637/jss.v074.i02

KLEIN, R. Utilização da teoria da resposta ao item no Sistema Nacional de Avaliação da Educação Básica (SAEB). *Ensaio: Avaliação e Políticas Públicas em Educação*, Rio de Janeiro v. 11, n. 40, p. 283-296, jul. 2003.

KLEIN, R. Utilização da teoria de resposta ao item no Sistema Nacional de Avaliação da Educação Basica (SAEB). *Revista Meta: Avaliação*, Rio de Janeiro, v. 1, n. 2, p. 125-140, sep. 2009.

KHORRAMDEL, L.; SHIN, H. J.; VON DAVIER, M. GDM software mdltm including parallel EM algorithm. *In*: DAVIER, M.; LEE, Y. S. (eds.). *Handbook of psychometric models for cognitive diagnosis*. [S. l.]: Springer, 2019. p. 603-628.

MAYDEU-OLIVARES, A. Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, [s. l.], v. 11, n. 3, p. 71-101, 2013. https://doi.org/10.1080/15366367.2013.831680

MCKINLEY, R. L.; MILLS, C. N. A Comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, [s. l.], v. 9, n. 1, p. 49-57, mar. 1985. https://doi.org/10.1177/014662168500900l0

OCDE. *PISA 2015 Technical Report*. Paris: OCDE, 2017.

OCDE. *PISA 2018 Technical Report*. Paris: OCDE, 2022.

OCDE. *Technical Report of the Survey of Adult Skills (PIAAC)*. 3rd ed. Paris: OECD, 2019.

OKUBO, T. *et al*. *PISA-based test for schools*: international linking study 2020. Paris: OECD, 2021. (OECD Education Working Paper, v. 244).

OLIVERI, M. E.; VON DAVIER, M. Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, v. 53, n. 3, p. 315-333, Sep. 2011.

R CORE TEAM. *R*: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2020.

CESGRANRIO. Relatório técnico pedagógico do SAEB 2007. Rio de Janeiro: Cesgranrio, 2008.

CESGRANRIO. Relatório técnico pedagógico do ENEM 2016. Rio de Janeiro: Cesgranrio, 2017.

ROBITZSCH, A. *et al*. *CDM*: cognitive diagnosis modeling. [S. l: s. n., 2020]. Disponível em: https://CRAN.R-project.org/package=CDM. Acesso em: 12 ago. 2024.

ROBITZSCH A, KIEFER T, WU M. *TAM*: test analysis modules. R package version 4.2-21. [S. n.: s.l., 2024. Disponível em: https://CRAN.R-project.org/package=TAM. Acesso em 12 ago. 2024

STEIGER, J. H. Notes on the Steiger–Lind (1980) Handout. *Structural Equation Modeling: A Multidisciplinary Journal*, [s. l.], v. 23, n. 6, p. 777-781, Sep. 2016. https://doi.org/10.1080/10705511.2016.1217487

STEIGER, J. H.; LIND, J. C. Statistically-based tests for the number of common factors. *In*: ANNUAL MEETING OF THE PSYCHOMETRIC SOCIETY, Iowa City, 1980. [S. n. t.].

TENNANT, A.; PALLANT, J. F. The root mean square error of approximation (RMSEA) as a supplementary statistic to determine fit to the Rasch model with large sample sizes. *Rasch Measurement Transactions*, [s. l.], v. 25, n. 4, p. 1348-1349, 2012.

WRIGHT, B. D.; MASTERS, G. N. *Rating scale analysis*. Chicago: Mesa Press, 1982.

YAMAMOTO, K.; KHORRAMDEL, L.; VON DAVIER, M. Scaling PIAAC cognitive data. *In*: OECD (ed.). *Technical report of the survey of adults skills (PIAAC)*. Paris: OECD, 2013. Chapter 17.

YEN, W. M. Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, [s. l.], v. 5, n. 2, p. 245-262, Apr. 1981. https://doi.org/10.1177/014662168100500212

YEN, W. M. Effects of Local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, v. 8, n. 2, p. 125-145, Apr. 1984. https://doi.org/10.1177/014662168400800201

CC BY

Informações sobre o autor

Ruben Klein: Ph.D. em Matemática pelo *Massachusetts Institute of Technology* (MIT), EUA. Consultor em Est'atística e Avaliação e Educacional da Fundação Cesgranrio. Membro da Associação Brasileira de Avaliação Educacional. Contato: ruben@cesgranrio.org.br

Dados: O conjunto de dados que dá suporte aos resultados deste estudo não está disponível publicamente, pois são provenientes de análises realizadas na Fundação Cesgranrio.

Conflitos de interesse: O autor declara que não possui nenhum interesse comercial, ou associativo que represente conflito de interesse em relação ao manuscrito.

Agradecimento: O autor agradece a Leandro Marino por comentários em uma versão preliminar desse artigo.