ARTIGO

Calibração com parâmetros de itens fixos

Ruben Klein [a]
Thales Akira Ricarte [b]

Resumo

O foco deste artigo é um método de calibração com parâmetros de itens fixos, *fixed item parameter estimation* (FIPC) para um modelo da teoria de resposta ao item (TRI). O artigo mostra como utilizar os pacotes mirt e irtplay do R e por simulação mostra que os resultados são equivalentes ao uso do método de grupos múltiplos, sendo um dos grupos o grupo de origem dos itens comuns com todos os parâmetros conhecidos.

Palavras-chave: FIPC. TRI. Mirt. Irtplay. Grupos Múltiplos.

1 Introdução

Na avaliação é necessário construir bancos de itens calibrados na mesma escala e, também, colocar diferentes testes na escala de um teste padrão original. Normalmente isso é feito se utilizando itens comuns entre diferentes testes. Alguns métodos são descritos em Kolen e Brennan (2014). Nesse artigo, trata-se somente de equalização para itens calibrados pela Teoria de Resposta ao item (TRI).

Talvez o método mais antigo seja o chamado calibração separada (*separated calibration*), no qual dois testes com itens comuns são calibrados separadamente e depois um deles é colocado na escala do outro. Esse método também é descrito em manuais de alguns *softwares* como o Bilog 3, Mislevy e Bock (1990) em sua versão DOS e atualmente em sua versão BilogMG para Windows (Toit, 2003). Na prática, esse método apresenta alguns problemas como os itens comuns não terem exatamente os mesmos parâmetros após a transformação de escala. Esse problema é resolvido fixando-se os parâmetros do primeiro grupo, chamado

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

Recebido em: 04 nov. 2024
Aceito em: 17 fev. 2025

de grupo de referência. Nos modelos logísticos de três parâmetros (3PL) na TRI, o parâmetro "c" de "*guessing*", apresenta mais problemas, pois não há transformação para ele e por isso costuma-se fixar o parâmetro "c" dos itens comuns na calibração do segundo grupo.

Outro método utilizado é o de grupos múltiplos (Bock, Zimowski, 1996). Neste método, dois ou mais grupos são calibrados simultaneamente, tendo os itens comuns entre os grupos uma calibração única, a melhor calibração conjunta, não a melhor calibração para cada grupo. Nesse método, um dos grupos é escolhido como o grupo de referência e, usualmente, fixa-se sua média como zero e desvio-padrão como um. As médias e desvios-padrão dos demais grupos são estimadas junto com os parâmetros. Esse método é implementado, por exemplo, no BilogMG e no *software* livre mirt (Chalmers, 2012).

Klein (2003) mostra como o método de grupos múltiplos foi utilizado no Sistema de Avaliação da Educação Básica (Saeb), 1997, e como foi adaptado para os Saeb seguintes. Esse método tem sido utilizado também no Exame Nacional do Ensino Médio (Enem) desde 2009.

Na primeira adaptação, tem-se um grupo ou mais já calibrados na mesma escala e se acrescentam os dados de um ou mais novos grupos com itens comuns e itens novos. O(s) grupo(s) antigo(s) tem que conter o grupo de referência, cuja média e desvio-padrão foram fixados em zero e um, respectivamente.

Na versão do BilogMG antiga para DOS, os parâmetros eram "fixados" pelas prioris e recalculados, ficando próximos dos valores fixados. Na versão para Windows, o parâmetro "c" no 3PL é fixado e os outros dois parâmetros reestimados ficando muito próximos dos valores fixados. Nesta versão, indica-se somente que os parâmetros foram fixados. Nas análises, posteriormente, pode-se manter os valores fixados inicialmente, em vez dos recalculados. No mirt, os valores fixados não mudam.

Na segunda adaptação, toma-se como grupo de referência um grupo que não era de referência e que na estimação passou a ter média e desvio-padrão estimado pelo método de grupos múltiplos. No BilogMG, esses valores são dados no arquivo .PH2 e no mirt pela função coef. Como descrito em Klein (2003; 2009), utiliza-se esses valores para transformar a média e o desvio-padrão desse grupo para zero e um, aplicando essa transformação aos parâmetros dos itens desse grupo. Após a calibração, aplica-se a transformação inversa para retornar a escala original.

No entanto, é interessante poder utilizar um método de equalização que utiliza somente os parâmetros dos itens. Nesse caso, seria fácil utilizar itens fixos vindos de um mesmo banco proveniente de aplicações diferentes, todos na mesma escala. Nesses métodos, a indeterminação dos modelos é resolvida pelos parâmetros dos itens fixos.

Kim (2006) compara, por simulação, cinco métodos de calibração com parâmetros de itens fixos (*fixed item calibration parameter* – FIPC) e mostra que somente um deles, o *multiple weights updating – multiple* EM *cycles* (MWU-MEM) apresenta bons resultados em todos os casos de simulação considerados. Kim menciona que o manual do BilogMG cita um procedimento de calibração com parâmetros fixos (FPC) e, também, como utilizar o Parscale (Toit, 2003). Segundo Kim, o procedimento do Parscale parece ser equivalente ao MWU-MEM e o procedimento do BilogMG ao NWU_MEM (*no prior weights updating – multiple* EM *cycles*) procedimento. Kim comparou esses dois procedimentos na simulação também e os resultados foram semelhantes aos do NWU-MEM e MWU-MEM.

Kang and Petersen (2009) fizeram um estudo de simulação comparando os métodos de calibração separada, *concurrent calibration* (que é o método dos grupos múltiplos), o procedimento FPC do BilogMG e o do Parscale. Os resultados foram que somente o procedimento do Parscale foi equivalente ao da calibração separada e ao dos grupos múltiplos.

Hoje em dia, existem *softwares* livres para efetuar essas análises. Citamos dois.

O pacote mirt com a função mirt que, se fixando os parâmetros dos itens e se configurando para estimar a média e o desvio-padrão do grupo, fornece resultados equivalentes aos do método com grupos múltiplos.

O *software* irtplay (Lim, Wells, 2020; 2022) foi elaborado para o R com o objetivo de implementar o método MWU-MEM, que o programa chama de MEM. Porém, atualmente essa biblioteca não está mais disponível no R. Este programa trabalha com somente 1 grupo.

Todos esses programas têm utilizado o método EM de otimização.

Este método FIPC pode ser estendido para vários grupos e a função multipleGroup do mirt fornece, novamente, resultados equivalentes ao uso de múltiplos grupos.

Destaca-se que a equalização dos itens do Programa Internacional de Avaliação de Estudantes (Pisa) para escolas com a escala do Pisa foi feita por esse método (Okubo *et al*., 2021). Okubo *et al*. (2021) mostram as equações de verossimilhança, inclusive para vários grupos que o artigo chama de modelo de misturas finitas, que é a análise de grupos múltiplos. O desenvolvimento de estimação é semelhante ao dos grupos múltiplos, só que, se fixando os parâmetros de alguns itens, se estimam as médias e desvios-padrão de todos os grupos.

Na próxima seção será exibida um estudo de simulação com esses métodos, o uso do mirt e do irtplay.

2 Estudo de simulação

Neste artigo, um estudo de simulação foi realizado para comparar a eficiência da recuperação dos parâmetros/curva dos itens em relação a sete casos. Para isso, foram simulados três populações ou grupos com distribuições normais N (0,1) para o grupo 1, N(0.5,1.2^2) para o grupo 2 e N(1,1.4^2) para o grupo 3, como em Kim (2006). Foram considerados 5.000 indivíduos para cada grupo. Cem itens foram simulados seguindo o 3PL no modo logístico, os parâmetros dos itens foram gerados como descrito a seguir: os parâmetros de discriminação "a" foram gerados a partir de uma distribuição uniforme U(0.5,3) , os parâmetros de dificuldade (ou posição) 'b' a partir de uma distribuição normal N(0,1) e o parâmetro "c" da assíntota inferior a partir de uma distribuição uniforme ((0.05,0.4). Foram feitas 50 replicações para cada caso.

Os indivíduos de cada grupo respondem a 40 itens, sendo 10 itens comuns entre os 1º e 2º grupos e 10 entre os 2º e 3º grupos, exceto no primeiro caso (os casos serão descritos a seguir) no qual todos os indivíduos responderam a todos os 100 itens. Nos casos de vários grupos, foram feitas simulações utilizando o BilogMG e mirt. Os casos considerados foram os seguintes:

Caso a. Três grupos, todos respondem a todos os itens. Estimação simultânea dos parâmetros dos itens.

Caso b. Três grupos, indivíduos respondem 40 itens por grupo, 10 itens em comum entre grupos. Estimação simultânea dos parâmetros dos itens.

Caso c. Três grupos, indivíduos respondem 40 itens por grupo, 10 itens em comum entre grupos. Primeiramente, apenas os parâmetros para o grupo 1 foram

calibrados. Fixando os parâmetros dos itens do grupo 1, estima-se os parâmetros dos itens novos dos grupos 2 e 3.

Caso d. Fixando os parâmetros dos itens do grupo 1, estima-se os parâmetros dos itens novos do grupo 2. Fixando os parâmetros dos itens dos grupos 1 e 2, estima-se os parâmetros dos itens novos do grupo 3.

Caso e. Fixando os parâmetros dos itens do grupo 1, estima-se os parâmetros dos itens novos do Grupo 2. Transforma-se os parâmetros dos itens do grupo 2 para que este tenha média 0 e desvio-padrão 1. Usando o grupo 2 transformado como referência, estima-se os parâmetros dos itens novos do grupo 3. Aplica-se a transformação inversa para obtenção dos parâmetros na escala (Klein, 2003).

Caso f. Fixa-se os parâmetros dos itens em comum dos grupos 1 e 2. Aplica-se o método FIPC para estimar os parâmetros dos itens novos do grupo 2. Procedimento análogo para os grupos 2 e 3, isto é, fixa-se os parâmetros dos itens em comum dos grupos 2 e 3. Aplica-se o método FIPC para estimar os parâmetros dos itens novos do grupo 3.

Caso g. Estima-se os parâmetros dos itens dos três grupos separadamente e utiliza-se os itens comuns para equalizar os parâmetros. Os parâmetros "c" dos itens comuns são fixados pelos resultados dos grupos 1 e posteriormente pelo grupo 2.

A simulação foi feita utilizando-se os *softwares* BilogMG, mirt e irtplay. As prioris são as utilizadas como no BilogMG, com o "c" centrado em 0.2. Para o mirt que utiliza a transformação do parâmetro "c", foi utilizado a distribuição normal com média log(0.2/(1-0.2)) = -1.386 e desvio-padrão 0.5. Esta distribuição sugerida pelo Prof. Pedro Barbetta dá resultados muito próximos dos resultados do BilogMG. A Tabela 1 apresenta as *prioris* utilizadas.

Tabela 1 - Distribuições *a priori* utilizadas

Software	**A**	**c**	**Θ**
BilogMG	lnorm(log(1.7),0.5)	beta(5,17)	N(0,1)
mirt	lnorm(log(1.7),0.5)	log(c/(1-c)) ~ N(-1.386,0.5)	N(0,1)
irtplay	lnorm(log(1.7),0.5)	beta(5,17)	N(0,1)

Fonte: Elaboração própria (2023)

A Tabela 2 apresenta o viés e o *root mean square error* (RMSE) para as estimativas dos três parâmetros. Os cinco primeiros casos de uso dos múltiplos grupos foram feitos com o BilogMG e com o mirt. O caso f com o FIPC só pôde ser feito com o mirt e o irtplay. Finalmente, o caso g, das calibrações separadas foi feito também somente com o mirt e o irtplay. Nesse último caso, como descrito, os parâmetros "c" dos itens comuns foram fixados após a calibração do grupo 1 e após a calibração do grupo 2. Como pode ser visto, os resultados são muito parecidos e bons.

Tabela 2 - Viés e RMSE dos parâmetros dos itens estimados para Casos "a" à "g"

Caso	*Software*	a		b		c	
		Viés	RMSE	Viés	RMSE	Viés	RMSE
A	BilogMG	-0.01	0.08	0.04	0.11	0.00	0.03
	mirt	-0.02	0.08	-0.01	0.13	0.00	0.04
B	BilogMG	0.00	0.15	0.03	0.15	0.00	0.05
	mirt	-0.05	0.17	-0.01	0.18	-0.01	0.06
C	BilogMG	0.07	0.18	0.04	0.15	0.00	0.05
	mirt	-0.01	0.18	0.00	0.17	0.00	0.05
D	BilogMG	0.05	0.18	0.05	0.15	0.00	0.05
	mirt	0.03	0.17	0.03	0.15	0.00	0.05
E	BilogMG	0.03	0.16	0.04	0.15	0.00	0.05
	mirt	0.03	0.16	0.03	0.15	0.00	0.05
F	mirt	0.02	0.18	0.02	0.16	0.00	0.05
	irtplay	0.06	0.21	0.04	0.15	0.00	0.05
G	mirt	0.02	0.19	0.02	0.17	0.00	0.05
	irtplay	0.06	0.22	0.05	0.17	0.01	0.05

Fonte: Elaboração própria (2023)

Tabela 3 - Médias e desvios-padrão das proficiências estimadas para Casos "a" a "g"

Caso	*Software*	Grupo 1		Grupo 2		Grupo 3	
		Média	DP	Média	DP	Média	DP
A	BilogMG	0.00	0.98	0.56	1.17	1.07	1.35
	mirt	-0.03	0.98	0.53	1.18	1.03	1.33

Continua

Continuação

B	BilogMG	0.00	0.95	0.56	1.14	1.08	1.30
	mirt	-0.01	0.96	0.55	1.17	1.05	1.32
C	BilogMG	0.00	0.95	0.55	1.10	1.01	1.19
	mirt	0.00	0.95	0.55	1.15	1.03	1.30
D	BilogMG	0.00	0.95	0.56	1.12	1.06	1.21
	mirt	0.00	0.95	0.55	1.14	1.03	1.28
E	BilogMG	0.00	0.95	0.56	1.12	1.06	1.26
	mirt	0.00	0.95	0.55	1.14	1.03	1.28
F	mirt	0.00	0.95	0.55	1.14	1.03	1.27
	irtplay	0.00	0.94	0.55	1.12	1.03	1.23
G	mirt	0.00	0.95	0.49	1.08	0.88	1.12
	irtplay	0.00	0.94	0.49	1.08	0.87	1.12

Fonte: Elaboração própria (2023)

A Tabela 3 mostra as médias das médias e desvios-padrão das proficiências estimadas em cada rodada da simulação. As proficiências foram calculadas pelo método *expected a posteriori* (EAP) e com *a priori* dada pela distribuição normal com a média e desvio-padrão do grupo dadas pelos *softwares*.

Observa-se que o desvio-padrão das proficiências é sempre menor que o do grupo, pois a variância do grupo é igual à soma da variância da esperança condicional da *posteriori* mais a esperança da variância condicional da *posteriori*. A esperança condicional da *posteriori* é a proficiência calculada pelo EAP. Percebe-se que a recuperação foi muito boa e muito parecidas em todos os casos.

Observa-se que, no caso g, de calibrações separadas, as médias e desvios-padrão estão um pouco abaixo dos outros casos, pois não há uma média e desvio-padrão de grupo para os grupos 2 e 3, utilizou-se *a priori* N(0,1). Talvez fosse o caso de usar os parâmetros de transformação para a média e desvio-padrão.

A Tabela 4 mostra o tempo de execução da simulação para os diversos casos com o mirt e o irtplay. Observa-se que o caso ‘f’, método FIPC pelo mirt, foi o mais rápido.

Tabela 4 - Tempo de execução dos casos "a" a "g"

Caso	Fonte	tempo (minutos)
A	mirt	88
B	mirt	87
C	mirt	30
D	mirt	29
E	mirt	20
F	mirt	8
	irtplay	24
G	mirt	14
	irtplay	22

Fonte: Elaboração própria (2023)

3 Distância entre curvas

Pode ser interessante calcular a distância entre duas curvas como no caso de duas curvas características de um item serem calibradas por dois *softwares* ou métodos diferentes para se tomar uma decisão se os dois métodos tiveram estimativas semelhantes. Não se deve comparar somente os parâmetros dos itens, mas sim suas curvas e focar especialmente na região onde as populações estão concentradas, como, por exemplo, entre os quantis 5% e 95%.

Pode-se tratar esse problema como no caso do DIF entre dois grupos (Klein, 2025), onde as duas curvas características substituem as "curvas" das proporções esperadas ou empíricas.

Como no DIF podemos definir:

MaxAdif = máximo das diferenças entre as curvas entre o máximo dos quantis 5% das proficiências dos dois grupos e o mínimo dos quantis 95% das proficiências dos dois grupos.

Na definição de RMSD, a diferença é entre as duas curvas e f(θ) é uma distribuição normal derivada das distribuições finais dos grupos nas duas calibrações. RMSD é definido por:

$$\textit{diferença entre curvas}: RMSD = \sqrt{(\int (P_o(\theta) - P_e(\theta)^2 \ f(\theta) d(\theta)}$$

Onde $(P_o(\theta) - P_e(\theta))^2$ é a diferença entre as curvas no ponto θ e $f(\theta)$ é uma distribuição.

Discretizando temos:

RMSD = Raiz quadrada (Soma (peso * (curva1 – curva2)^2)),

onde o peso é definido por $f(\theta)d(\theta)$, nos pontos de quadratura utilizados.

Este peso pode ser obtido pelas distribuições finais das duas calibrações ou pelas proporções empíricas, como no caso de Dif. Klein (2025) sugere pegar pesos proporcionais a raiz quadrada do produto dos pesos ou proporcionais ao produto dos pesos. Caso as distribuições finais obtidas pelos dois métodos sejam muito próximas, pode-se pegar somente uma das distribuições. Por exemplo, a raiz quadrada do produto dos pesos reproduz o peso se esses forem iguais. Nos exemplos da próxima seção, essas estatísticas são calculadas.

4 Exemplos com dados reais

Nesta seção, apresentamos dois exemplos com dados reais, o primeiro com somente um grupo para calibrar e o segundo com três grupos. Os resultados do uso do mirt e irtplay são comparados aos resultados obtidos com o BilogMG.

4.1 Exemplo 1

O primeiro exemplo é o de uma aplicação feita com o método dos grupos a partir do BilogMG e há somente um grupo para se calibrar. Os três primeiros grupos tinham parâmetros conhecidos e o objetivo era calibrar o quarto grupo, que tinha itens comuns e itens novos. O primeiro grupo tinha sofrido uma transformação para que sua média e desvio-padrão fossem 0 e 1, respectivamente. Após a calibração, todos os parâmetros dos itens foram transformados pela transformação inversa. Nesse caso não houve necessidade de abandonar itens ou considerar itens novos.

Fixando-se os parâmetros dos itens comuns no 4º grupo na escala final, aplicou-se o método FIPC utilizando-se o mirt e o irtplay. Neste caso de itens bem-comportados, os resultados foram quase iguais, inclusive na estimativa dos parâmetros novos, como mostram as Tabelas 5 a 7.

A Tabela 5 mostra um resumo das diferenças entre os parâmetros dos itens estimados pelo BilogMG e mirt e entre o BilogMG e o irtplay. A Tabela 6 mostra um resumo das diferenças entre as proficiências estimadas pelo BilogMG e pelos mirt e irtplay. A proficiência transformada foi para uma escala com média 250 e desvio-padrão 50. Finalmente, a Tabela 7 mostra as correlações de Fisher e de Spearman entre as proficiências estimadas pelos três *softwares*.

Tabela 5 - Diferenças dos parâmetros dos itens estimados na aplicação entre o BilogMG e o mirt (irtplay)

Software	**Resumo**	**BilogMG**					
		a		**b**		**c**	
		difer.	**dist.**	**difer.**	**dist.**	**difer.**	**dist.**
Mirt	Mínimo	-0.08	0.00	-0.01	0.00	-0.01	0.00
	1º quartil	0.02	0.03	0.00	0.00	0.00	0.00
	Mediana	0.04	0.04	0.00	0.01	0.01	0.01
	3º quartil	0.05	0.06	0.03	0.03	0.03	0.03
	Máximo	0.13	0.13	0.19	0.19	0.04	0.04
Irtplay	Mínimo	0.09	0.09	-0.01	0.00	0.00	0.00
	1º quartil	0.11	0.11	0.00	0.01	0.01	0.01
	Mediana	0.17	0.17	0.03	0.03	0.02	0.02
	3º quartil	0.21	0.21	0.08	0.08	0.04	0.04
	Máximo	0.25	0.25	0.21	0.21	0.05	0.05

Fonte: Elaboração própria (2023)

Tabela 6 - Diferenças das proficiências estimadas na aplicação entre o BilogMG e o mirt (irtplay)

Software	**Resumo**	**BilogMG**			
		proficiência		**proficiência transformada**	
		diferença	**distância**	**diferença**	**distância**
Mirt	Mínimo	-0.08	0.00	-4.68	0.00
	1º quartil	-0.03	0.01	-1.62	0.51
	Mediana	0.00	0.03	-0.03	1.48
	3º quartil	0.02	0.05	1.34	2.55
	Máximo	0.07	0.08	3.99	4.68

Continua

Continuação

Irtplay	Mínimo	-0.20	0.00	-10.93	0.00
	1º quartil	-0.07	0.03	-3.72	1.68
	Mediana	0.00	0.07	-0.13	3.89
	3º quartil	0.07	0.10	4.04	5.71
	Máximo	0.17	0.20	9.60	10.93

Fonte: Elaboração própria (2023)

Tabela 7 - Correlações das proficiências estimadas entre os softwares na aplicação

***Software* 1**	***Software* 2**	**Correlação**	
		Fisher	**Spearman**
BilogMG	mirt	0.9978	0.9999
BilogMG	irtplay	0.9883	0.9995
mirt	irtplay	0.9951	0.9998

Fonte: Elaboração própria (2023)

A Tabela 8 mostra um resumo das distâncias entre as curvas obtidas pelo BilogMG e pelo mirt. Utilizou-se o peso derivado da análise do mirt. Como se pode ver, as curvas são praticamente iguais.

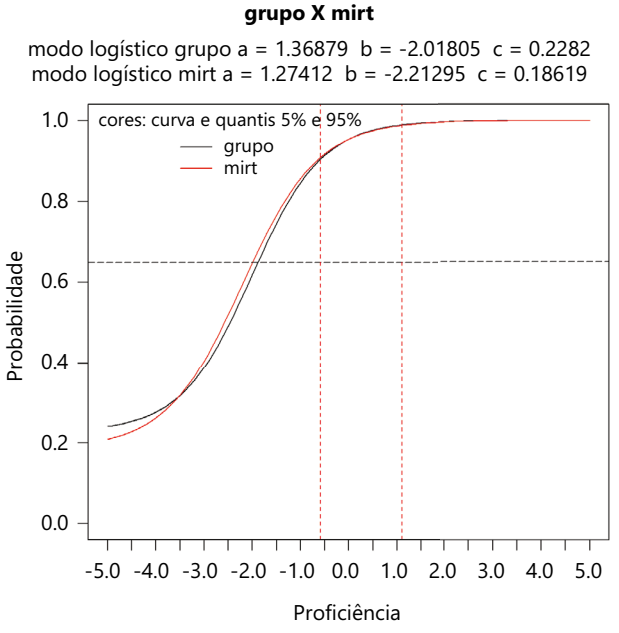
Tabela 8 - Resumo das diferenças das estatísticas MaxAdif e RMSD para as curvas obtidas pelo BilogMG e pelo mirt

Resumo	**Mínimo**	**1º quartil**	**Mediana**	**3º quartil**	**Máximo**
MaxAdif	0.000	0.000	0.003	0.004	0.009
RMSD	0.000	0.000	0.001	0.002	0.004

Fonte: Elaboração própria (2023)

A seguir, no Gráfico 1, exibe-se o item com a maior diferença absoluta no parâmetro "b", que tem um MaxAdif de 0.004 e um RMSD de 0.002. Observa-se que as curvas são praticamente iguais entre os quantis 5% e 95%.

Gráfico 1 - Gráfico das duas curvas do item com a maior diferença no parâmetro "b"



Fonte: Elaboração própria (2023)

4.2 Exemplo 2

O segundo exemplo tem como objetivo calibrar três grupos, com itens comuns entre o primeiro e segundo grupo e entre o segundo e terceiro grupo. A análise original, feita com o BilogMG, continha 12 grupos, os seis primeiros com todos os itens fixados e os seis últimos a serem calibrados. Havia itens comuns entre o 7º ao 9º grupo e entre os três últimos (10º ao 12º). Os três últimos grupos são os grupos que se quer calibrar e colocar na escala dos seis primeiros. É claro que havia itens comuns entre os seis primeiros grupos e os três últimos. Observa-se que já havia transformação dos primeiros seis grupos para que o grupo de referência tivesse média zero e desvio-padrão um. Essa análise de múltiplos grupos teve duas rodadas, tendo havido alguns itens abandonados e outros considerados novos por problemas de parâmetros, gráficos, ajuste e DIF. Nesse exemplo, considerou-se as decisões tomadas após a 1ª rodada.

Os três grupos continham 221 itens, sendo 109 com parâmetros fixos.

Se fossemos começar a análise com o método FIPC desde o começo, deveríamos fazer as mesmas análises de parâmetros, gráficos e ajuste conforme Klein (2025). Se tivéssemos os grupos com itens fixados com seus dados originais, poderíamos também fazer estudo de DIF.

Para essa análise, se utilizou a função multipleGroup do mirt, com os itens comuns com os seis primeiros grupos fixados.

A Tabela 9 mostra um resumo das diferenças entre os parâmetros dos itens estimados e a Tabela 10 mostra o resumo das diferenças entre as proficiências. Verifica-se que essas diferenças são bem pequenas. A proficiência foi transformada para uma escala com média 0 e desvio-padrão 1. A Tabela 11 mostra as correlações de Fisher e Spearman entre as proficiências estimadas para cada grupo. Pode-se ver que essas correlações são muito próximas de 1.

Tabela 9 - Diferenças dos parâmetros dos itens estimados na aplicação entre o BilogMG e o mirt para os três grupos

	Resumo	**a**	**B**	**c**
grupo 1	Mínimo	-0.022	-0.078	-0.077
	1º quartil	0.000	-0.010	-0.003
	Mediana	0.000	-0.001	-0.001
	3º quartil	0.011	0.000	0.000
	Máximo	0.038	0.005	0.001
grupo 2	Mínimo	-0.033	-0.144	-0.071
	1º quartil	-0.002	-0.013	-0.004
	Mediana	0.000	-0.004	0.000
	3º quartil	0.000	0.000	0.000
	Máximo	0.267	0.005	0.000
grupo 3	Mínimo	-0.033	-0.245	-0.080
	1º quartil	0.000	-0.013	-0.001
	Mediana	0.000	0.000	0.000
	3º quartil	0.005	0.000	0.000
	Máximo	0.174	0.000	0.000

Fonte: Elaboração própria (2023)

Tabela 10 - Diferenças entre as proficiências estimadas na aplicação entre o BilogMG e o mirt para os três grupos

	Resumo	proficiência	proficiência transformada
grupo 1	Mínimo	-0.027	-1.509
	1º quartil	-0.002	-0.093
	Mediana	0.001	0.042
	3º quartil	0.004	0.244
	Máximo	0.050	2.811
grupo 2	Mínimo	-0.035	-1.949
	1º quartil	-0.002	-0.098
	Mediana	0.000	0.001
	3º quartil	0.004	0.210
	Máximo	0.065	3.637
grupo 3	Mínimo	-0.130	-7.275
	1º quartil	-0.003	-0.140
	Mediana	0.000	0.008
	3º quartil	0.004	0.241
	Máximo	0.066	3.674

Fonte: Elaboração própria (2023)

Tabela 11 - Correlações entre as proficiências estimadas na aplicação entre o BilogMG e o mirt para os três grupos

	Fisher	Spearman
grupo 1	0.99998	0.99999
grupo 2	0.99998	0.99999
grupo 3	0.99997	0.99998

Fonte: Elaboração própria (2023)

Finalmente, a Tabela 12 mostra um resumo das distâncias entre as curvas obtidas pelo BilogMG e pelo mirt para cada grupo. Utilizou-se o peso derivado da análise do mirt. Como se pode ver as curvas são praticamente iguais.

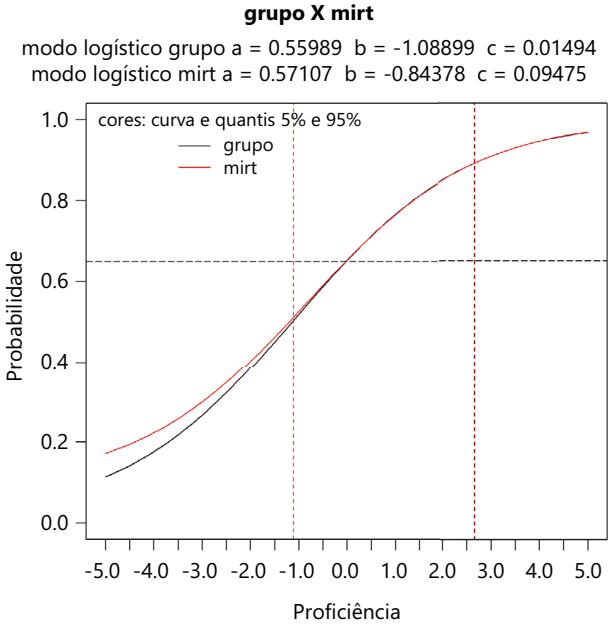
Tabela 12 - Resumo das diferenças das estatísticas MaxAdif e RMSD para as curvas obtidas pelo BilogMG e pelo mirt, para cada grupo

	Resumo	MaxAdif	RMSD
grupo 1	Mínimo	0.000	0.000
	1º quartil	0.000	0.000
	Mediana	0.002	0.001
	3º quartil	0.005	0.002
	Máximo	0.015	0.008
grupo 2	Mínimo	0.000	0.000
	1º quartil	0.000	0.000
	Mediana	0.002	0.001
	3º quartil	0.003	0.002
	Máximo	0.015	0.009
grupo 3	Mínimo	0.000	0.000
	1º quartil	0.000	0.000
	Mediana	0.000	0.000
	3º quartil	0.004	0.002
	Máximo	0.015	0.008

Fonte: Elaboração própria (2023)

A seguir, no Gráfico 2, exibe-se o item com a maior diferença absoluta no parâmetro "b" em todos os grupos. É um item do grupo 3. Esse item tem um MaxAdif de 0.009 e um RMSD de 0.005. Observa-se que as curvas são praticamente iguais entre os quantis 5% e 95%.

Gráfico 2 - Gráfico das duas curvas do item com a maior diferença no parâmetro "b" em todos os três grupos.



Fonte: Elaboração própria (2023)

5 Conclusão

O método FIPC mostrou ser uma boa alternativa ao uso dos métodos de grupos múltiplos quando há itens comuns com parâmetros fixados. É de uso simples e pelas simulações e pelos exemplos apresenta resultados muito próximos aos métodos de grupos múltiplos em suas várias formas. As simulações também mostram que para os métodos de grupos múltiplos pode-se usar o BilogMG ou o mirt com as *prioris* consideradas. A grande vantagem do FIPC é só requerer os parâmetros dos itens fixados e não ter necessidade de dados adicionais.

Fixed item parameter calibration

Abstract

This paper focus on a FIPC (Fixed item parameter estimation) of an IRT (Item Response Theory) model. This shows how to use the R packages mirt and irtplay and by simulation that its results are equivalent to the method of multiple groups, where one of the groups is the group of origin of the common items with all parameters known.

Keywords: *FIPC. TRI. Mirt. Irtplay. Multiple groups.*

Calibración con parámetros fijos

Resumen

El objetivo de este artículo es un método de calibración con parámetros de ítem fijos, FIPC (estimación de parámetros de ítem fijos) para un modelo IRT (teoría de respuesta al ítem). El artículo muestra cómo usar los paquetes de R mirt e irtplay y, mediante simulación, muestra que los resultados son equivalentes a usar el método de grupos múltiples, siendo uno de los grupos el grupo de origen de los elementos comunes con todos los parámetros conocidos.

Palabras clave: *FIPC. TRI. Mirt. Irtplay. Grupos Múltiples.*

Referências

BOCK, R. D.; ZIMOWSKI, M. F. Multiple group IRT. *In*: LINDEN, W. J.; HAMBLETON, R. K. (Eds.). *Handbook of modern item response theory*. New York: Springer, 1996. p. 433-448.

CHALMERS, R. P. Mirt: a multidimensional item response theory package for the R environment. *Journal of Statistical Software*, Innsbruck, v. 48, n. 6, maio 2012. https://doi.org/10.18637/jss.v048.i06

KANG, T.; PETERSEN, N. S. Linking item parameters to a base scale. *ACT Research Report Series*, ago 2009.

KIM, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurament*, [S. l.], v. 43, n. 1, p. 355-381, 2006. https://doi.org/10.1111/j.1745-3984.2006.00021.x

KLEIN, R. Medidas de qualidade de ajuste e de DIF (differential item functioning). *Ensaio: Avaliação e Políticas Públicas em Educação*, Rio de Janeiro, v. 33, n. 126, p. 1-25, jan. 2025. https://doi.org/10.1590/S0104-40362025003305142

KLEIN, R. Utilização da teoria de 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./set. 2003.

KOLEN, M.; BRENNAN, R. *Test equating, scaling, and linking*. 3. ed. New York: Springer, 2014.

LIM, H; WELLS, C. S. [RETRACTED ARTICLE] irtplay: an R package for unidimensional item response theory modeling. *Journal of Statistical Software*, [S. l.], v. 103, n. 12, p. 1-42, 2022. https://doi.org/10.18637/jss.v103.i12

LIM, H; WELLS, C. S. Irtplay: an R package for online item calibration, scoring, evaluation of model fit, and useful functions for unidimensional IRT. *Applied Psychological Measurement*, [S. l.], v. 44, n. 7-8, p. 563-565, out. 2020. https://doi.org/10.1177/0146621620921247

MISLEVY, R. J.; BOCK, R. D. *Bilog 3*: item analysis and test scoring with binary logistic models. 2. ed. [S. l.]: Scientific Software International, 1990.

OKUBO, T. *et al*. *PISA Based test for schools*: international linking study 2020. OECD Education Working Paper No. 244. Paris: Organisation de Coopération et de Développement Économiques, 2021. https://doi.org/10.1787/ef1356ae-en

TOIT, M. (Ed.). *IRT from SSI*: bilogmg, multilog, parscale, testfact. Lincolnwood: Scientific Software International, 2003.

Informações sobre os autores

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

Thales Akira Ricarte: Doutor em Estatística pela Universidade de São Paulo. Pesquisador da Fundação Cesgranrio. Contato: thalesamr@gmail.com

Contribuição dos autores: Ruben Klein e Thales Akira Ricarte – coleta e análise dos dados da pesquisa, discussão dos resultados, concepção, elaboração e escrita do manuscrito, revisão de versões e revisão crítica do conteúdo.

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: Os autores declaram que não possuem nenhum interesse comercial ou associativo que represente conflito de interesse em relação ao manuscrito.

Editores que avaliaram este artigo

Fátima Cunha
Érika Dias

Ruben Klein, Thales Akira Ricarte
Calibração com parâmetros de itens fixos
Fixed item parameter calibration
Calibración con parámetros fijos

Ensaio: Avaliação e Políticas Públicas em Educação
vol. 33, núm. 127, e0255144, 2025
Fundação CESGRANRIO,
ISSN: 0104-4036
ISSN-E: 1809-4465

DOI: https://doi.org/10.1590/S0104-40362025003305144