Bilog-MG x *Mirt* na Prova São Paulo 2024: uma comparação necessária em tempos de linguagem aberta e comunidade ativa

Resumo

A Prova São Paulo 2024, conduzida pelo Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos (Cebraspe), em parceria com a Secretaria Municipal de Educação de São Paulo (SME-SP), utilizou a Teoria da Resposta ao Item (TRI), implementada no *software* proprietário Bilog-MG para avaliar o aprendizado dos estudantes da rede municipal. A análise do desempenho mostrou um aumento geral das médias ao longo dos anos escolares, com destaque para o ciclo de alfabetização. Visando investigar a viabilidade de ferramentas de código aberto, as análises foram conduzidas em paralelo com o pacote *mirt*, da linguagem R, comparando-se as medidas obtidas por ambos os *softwares*. A forte correlação nos resultados da maioria dos itens sugere o potencial do *mirt* como alternativa ao Bilog-MG. A iniciativa representa um avanço para o uso de ferramentas de análise mais flexíveis no monitoramento da educação municipal.

Palavras-chave: Prova São Paulo, Teoria da Resposta ao Item, Bilog-MG e *Mirt.*

As Medidas da Prova São Paulo 2024 estimadas por meio do Bilog-MG

Ediane Nascimento Ferreira[1]
Coordenação de Estatística do Cebraspe
Brasília, Distrito Federal, Brasil
ediane.ferreira@cebraspe.org.br

Thiago Fernando Ferreira Costa[2]
Divisão de Avaliação – SME-SP
São Paulo, São Paulo, Brasil
thiagocosta@sme.prefeitura.sp.gov.br

Marcus Vinícius Araújo Soares[3]
Diretoria de Avaliação e Educação do Cebraspe
Brasília, Distrito Federal, Brasil
mv@cebraspe.org.br

Luciana Carvalho Ramos[4]
Coordenação de Ensino Pesquisa e Avaliação do Cebraspe
Brasília, Distrito Federal, Brasil
luciana.ramos@cebraspe.org.br

Priscila Trogo Pereira[5]
Coordenação de Ensino Pesquisa e Avaliação do Cebraspe
Brasília, Distrito Federal, Brasil
priscila.pereira@cebraspe.org.br

Resumo

Este artigo apresenta a análise estatística das medidas da Prova São Paulo 2024, utilizando a Teoria da Resposta ao Item (TRI) com o modelo de três parâmetros de Birnbaum, implementado no *software* Bilog-MG. A avaliação do aprendizado na rede municipal, conduzida pelo Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos (Cebraspe), em parceria com a Secretaria Municipal de Educação de São Paulo (SME-SP), envolveu um número expressivo de estudantes e utilizou uma estruturação rigorosa e análise psicométrica robusta para garantir a validade das medidas, cujas estimativas foram comparadas com as obtidas pelo pacote *mirt*, da linguagem R. A análise revelou um aumento geral da proficiência média ao longo dos anos escolares, com destaque para o crescimento no ciclo de alfabetização, o que fornece subsídios para o planejamento estratégico da SME-SP, das Diretorias Regionais de Educação (DREs) e das escolas, visando aprimorar a qualidade da educação.

Palavras-chave: Prova São Paulo; Teoria da Resposta ao Item; Medidas, Bilog-MG; *Mirt.*

[1] O autor 1 agradece às pessoas que o ajudaram.
[2] O autor 2 agradece às pessoas que o ajudaram.
[3] O autor 3 agradece às pessoas que o ajudaram.
[4] O autor 4 agradece às pessoas que o ajudaram.
[5] O autor 5 agradece às pessoas que o ajudaram.

1 Introdução

A Prova São Paulo, implementada em 2007 pela Secretaria Municipal de Educação de São Paulo (SME-SP), consolidou-se como um instrumento essencial de monitoramento da educação municipal, avaliando o rendimento escolar dos estudantes em Língua Portuguesa, Matemática, Ciências Naturais, Ciências Humanas e Produção de Texto. Ao longo dos anos, essa avaliação tem passado por aprimoramentos contínuos, visando oferecer um diagnóstico abrangente do seu progresso na rede. Um desses avanços foi a introdução da Provinha São Paulo em 2017. O objetivo central das duas avaliações, Provinha e Prova São Paulo, é fornecer informações sistematizadas sobre o aprendizado dos estudantes da rede municipal, dados que, ao serem analisados e utilizados pelos profissionais da educação, constituem um importante subsídio para o planejamento estratégico da própria Secretaria, das Diretorias Regionais de Educação (DREs) e das equipes escolares.

Em 2024, o Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos (Cebraspe), em parceria técnica estratégica com a SME-SP, foi responsável pela concepção, organização e execução da Provinha e da Prova São Paulo, cujas análises dos resultados também se deu em colaboração com a equipe técnica da Secretaria. Essas avaliações alcançaram um universo de 385.809 estudantes, distribuídos em 603 unidades educacionais da rede municipal, abrangendo desde o 2º ano do Ensino Fundamental até a 3ª série do Ensino Médio regular, além dos estudantes da Educação de Jovens e Adultos em suas diversas modalidades. A participação equitativa de estudantes surdos e cegos foi garantida por meio de recursos de acessibilidade, como itens audiodescritos e interpretação em Libras. A fim de possibilitar uma análise precisa e confiável dos resultados dessa abrangente avaliação, a estruturação dos testes foi uma etapa crucial.

Para garantir a validade e a confiabilidade dos dados coletados na edição de 2024, os testes da Provinha e da Prova São Paulo foram cuidadosamente estruturados, em consonância com o Currículo da Cidade. A técnica dos Blocos Incompletos Balanceados (BIB) foi empregada como estratégia eficiente para realizar uma avaliação abrangente e equitativa das habilidades nas diversas áreas de conhecimento avaliadas. Dada a complexidade e a riqueza dos dados provenientes dessas avaliações estruturadas com a técnica BIB, a análise psicométrica aprofundada, essencial para garantir a comparabilidade dos resultados ao longo do tempo e com as avaliações educacionais na escala do Sistema de Avaliação da Educação Básica (SAEB), demandou uma metodologia estatística robusta, como a Teoria da Resposta ao Item (TRI).

Este artigo, portanto, apresentará as medidas de desempenho dos estudantes da rede municipal de São Paulo em Língua Portuguesa e Matemática, do 2º ao 9º ano, e em Ciências Naturais, do 3º ao 9º ano, utilizando a TRI sob o modelo de três parâmetros de Birnbaum (Andrade et al., 2000; Baker e Kim, 2004), no *software* Bilog-MG, referência em análises de avaliações em larga escala. A validade dos resultados desta análise foi sustentada pela convergência, em média, com as estimativas obtidas pelo pacote *mirt* da linguagem R, conforme a verificação da equipe técnica da SME-SP.

2 Metodologia

O modelo de três parâmetros de Birnbaum (Andrade et al., 2000; Baker; Kim, 2004), no âmbito da TRI, foi adotado para a análise dos itens de múltipla escolha da Provinha e da Prova São Paulo 2024. Esse modelo estatístico é utilizado para avaliar traços latentes, representando a relação entre a probabilidade de resposta correta e a habilidade do avaliado (Andrade; Tavares; Valle, 2000). A escolha desse modelo possibilitou a equalização dos itens nas escalas SAEB, para Língua Portuguesa e Matemática, e na escala da própria Prova São Paulo, para Ciências Naturais. A calibração realizada com o *software* Bilog-MG estimou conjuntamente os parâmetros de dificuldade (*b*), discriminação (*a*) e probabilidade de acerto casual (*c*) desse modelo, a partir da distribuição dos padrões de resposta dos estudantes.

A equalização na escala SAEB ocorreu com o uso de bases clones, devido à ausência das bases de dados originais. A estimação dos parâmetros dos novos itens foi feita por máxima verossimilhança marginal, assumindo distribuição normal para as proficiências. A qualidade dos itens equalizados considerou critérios como correlação bisserial inferior a 0,01, problemas de convergência, estimativas atípicas e erros padrão elevados. Os itens que não atenderam a esses critérios foram excluídos da análise.

O Funcionamento Diferenciado do Item (DIF) foi investigado comparando itens comuns entre os grupos da equalização com o auxílio do Bilog-MG, focando na região entre os percentis 5 e 95. Itens com diferença de probabilidade de acerto superior a 0,15 foram considerados com DIF e recalibrados. A proficiência dos estudantes foi estimada por método análogo ao utilizado pelo SAEB, considerando o padrão de resposta e os parâmetros dos itens. A interpretação pedagógica dos resultados utilizou a escala de desempenho do SAEB, cujos níveis de proficiência foram definidos e validados por especialistas das áreas de conhecimento.

3 Resultados

A partir da análise, apresentam-se os resultados das áreas de conhecimento avaliadas na Prova e na Provinha São Paulo 2024, cujos dados estão organizados na Tabela 1. A análise concentra-se na evolução da proficiência dos estudantes do 2º ao 9º ano do Ensino Fundamental, considerando-se as médias de proficiência dos estudantes da rede municipal por área de conhecimento.

Na Tabela 1, observa-se em Língua Portuguesa um crescimento progressivo na média de proficiência ao longo dos anos escolares, com valores que variaram de 159,7, no 2º ano, a 229,7, no 9º ano. Os resultados do ciclo de alfabetização de 2024 revelaram avanços em relação à edição anterior: a média de proficiência do 2º ano cresceu 17,4 unidades, e a do 3º ano, 17,6 unidades na escala de proficiência. Houve também crescimento nas médias dos demais anos escolares, exceto no 7º ano (redução de 0,1 unidade na média de proficiência); contudo, esse crescimento não ultrapassou 7,6 unidades, registrado no 4º ano.

Os resultados de Matemática da rede municipal também indicaram progressão na proficiência média dos estudantes ao longo dos anos escolares, variando de 160,5, no 2º ano, a 229,3, no 9º ano. O ciclo de alfabetização demonstrou avanço significativo em relação à aplicação de 2023, com um aumento de 23,8 unidades na média de proficiência do 2º ano e de 12,5 unidades na do 3º ano. A trajetória de crescimento foi observada na maioria dos anos escolares, com exceção dos dois anos finais – 8º ano, com redução de 0,5 unidade, e 9º ano, com redução de 2,2 unidades. Entre os anos escolares que apresentaram avanços, o 4º ano alcançou o maior ganho (5,8 unidades).

Em Ciências Naturais, observou-se uma tendência semelhante, com a proficiência média variando de 159,0, no 3º ano, a 229,9, no 9º ano. O 3º ano apresentou a maior elevação em relação à edição de 2023 (12,5 unidades), enquanto o 9º ano registrou uma redução de 1,3 unidades na proficiência média.

Tabela 1: **Médias de proficiência em Língua Portuesa, Matemática e Ciências Naturais da rede municipal de São Paulo em 2024**

Ano escolar	Média de Proficiência (total de estudantes avaliados)		
	Língua Portuguesa	Matemática	Ciências Naturais
2º	159,7 (42.132)	160,5 (44.295)	
3º	169,0 (44.886)	166,1 (46.691)	159,0 (46.016)
4º	173,2 (44.617)	173,4 (46.425)	161,9 (45.595)
5º	188,2 (43.185)	190,6 (44.791)	186,1 (43.716)
6º	199,3 (39.070)	199,2 (40.411)	201,5 (39.663)
7º	207,6 (37.985)	210,1 (39.300)	213,4 (38.471)
8º	217,3 (41.083)	223,2 (42.619)	222,5 (41.419)
9º	229,7 (40.206)	229,3 (41.809)	229,9 (40.363)

Fonte: Cebraspe/SME-SP

4 Conclusões e Considerações Finais

A Prova e a Provinha São Paulo 2024, conduzidas pelo Cebraspe, em parceria com a SME-SP, forneceram um panorama abrangente do aprendizado na rede municipal, envolvendo muitos estudantes de diferentes perfis. As avaliações foram estruturadas com rigor, alinhadas ao Currículo da Cidade e à técnica BIB, e analisadas com metodologia psicométrica robusta (TRI), visando garantir a validade e a confiabilidade dos resultados. A análise da proficiência dos estudantes (do 2º ao 9º ano) em Língua Portuguesa, Matemática e Ciências Naturais revelou um aumento geral ao longo dos anos escolares, indicando desenvolvimento progressivo.

Um destaque importante foi o crescimento, em 2024, do desempenho no ciclo de alfabetização em relação à edição anterior, o que pode refletir o impacto positivo das ações de capacitação promovidas pela SME-SP em 2024. A validade dos resultados foi reforçada pela convergência com as estimativas obtidas pela equipe técnica da SME-SP, utilizando o pacote *mirt* da linguagem R, que, além de gerar resultados consistentes com a escala SAEB, oferece maior flexibilidade e acessibilidade em comparação com o *software* Bilog-MG. Esses resultados são valiosos para o planejamento estratégico da SME-SP, DREs e escolas.

5 Referências

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C. *Teoria da resposta ao item: conceitos e aplicações*. São Paulo: ABE - Associação Brasileira de Estatística, 2000.

BAKER, F. B.; KIM, S.H. *Item Response Theory – Parameter Estimation Techniques.* 2. ed. New York: Marcel Dekker, Inc., 2004.

As medidas da Prova São Paulo 2024 estimadas por meio do pacote *Mirt* da linguagem R

Thiago Fernando Ferreira Costa[1]

Divisão de Avaliação – SME-SP
São Paulo, São Paulo, Brasil
thiagocosta@sme.prefeitura.sp.gov.br

Ediane Nascimento Ferreira[2]

Coordenação de Estatística do Cebraspe
Brasília, Distrito Federal, Brasil
ediane.ferreira@cebraspe.org.br

Thais Barros de Paula Capel[3]

Divisão de Avaliação – SME-SP
São Paulo, São Paulo, Brasil
thais.capel@sme.prefeitura.sp.gov.br

Michelly Francini Brassaroto do Amaral[4]

Divisão de Avaliação – SME-SP
São Paulo, São Paulo, Brasil
Michelly.amaral@sme.prefeitura.sp.gov.br

Resumo

Este artigo tem o objetivo de apresentar os resultados das análises estatísticas realizadas no âmbito da Prova São Paulo (PSP) de 2024. A instituição legalmente responsável pela avaliação foi o Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos (Cebraspe), em parceria com a Divisão de Avaliação da Secretaria Municipal de Educação de São Paulo (DA/SME-SP). As análises estatísticas que envolveram a PSP nos últimos anos demandaram análises simultâneas entre a instituição vencedora para realizar a avaliação e o próprio setor técnico da Secretaria. Dessa maneira, em 2024, além dessas análises, dois modelos de *software* foram utilizados também de forma paralela. Sendo assim, o Cebraspe realizou as análises seguindo o modo tradicional, em que o Bilog-MG e todas as técnicas já amplamente divulgadas foram implementadas, e a Divisão de Avaliação (DA) utilizou o pacote *mirt* da linguagem R. Neste artigo, esses resultados serão apresentados para debate.

Palavras-chave: Teoria da Resposta ao Item; Bilog-MG; *Mirt*; Prova São Paulo.

[1] O autor 1 agradece às pessoas que o ajudaram.
[2] O autor 2 agradece às pessoas que o ajudaram.
[3] O autor 3 agradece às pessoas que o ajudaram.
[4] O autor 4 agradece às pessoas que o ajudaram.

1 Introdução

As análises da Prova São Paulo (PSP) adotam premissas semelhantes às do Sistema de Avaliação da Educação Básica (SAEB) para permitir a comparação entre os resultados, levando em consideração que as metodologias de aplicação e as definições de cálculo das médias são diferentes entre as provas. De todo modo, a PSP tem mostrado medidas coerentes com a evolução do SAEB ao longo dos anos, evidenciando que o progresso em uma prova geralmente acompanha o progresso na outra.

Assim como o SAEB, a PSP tradicionalmente empregou *software* proprietário, visando à estabilidade das medidas para garantir a comparabilidade com a escala do SAEB e evitar desconfianças dos usuários sobre essas medidas e resultados. A ocorrência de muitas mudanças drásticas de um ano letivo para outro impossibilitaria manter a explicação do ganho de proficiência para os profissionais das escolas.

De todo modo, com o avanço das linguagens abertas, principalmente daquelas mais utilizadas pela comunidade acadêmica, como o R, diversas funções foram construídas para implementar ferramentas estatísticas especializadas em processos descritivos e inferências. Nesse sentido, o pacote *mirt*, na linguagem R, foi elaborado para implementar diversos modelos da Teoria da Resposta ao Item (TRI).

Dessa maneira, este resumo expandido apresenta as análises realizadas com o pacote *mirt* sobre os dados da Prova São Paulo 2024 de Língua Portuguesa, Matemática e Ciências Naturais. Para viabilizá-las, foram empregadas duas metodologias: a primeira consistiu na elaboração de bases com respostas fictícias a partir dos parâmetros do modelo logístico de três parâmetros (3PL) – discriminação, relacionada à diferenciação de estudantes; dificuldade, que indica o nível de habilidade necessário para acertar o item; e probabilidade de acerto ao acaso por estudantes de baixa proficiência. A segunda metodologia, apresentada por Klein e Ricarte (2025), utilizou o método de calibração com elementos de itens fixos para o modelo 3PL da TRI. A lógica subjacente é que as estimativas geradas na primeira metodologia (utilizando o Bilog-MG) e na segunda metodologia (utilizando o *software* R) seriam comparadas.

2 Metodologia

A metodologia utilizada para analisar os construtos de Língua Portuguesa e Matemática da PSP 2024 é descrita em Klein (2003), método de grupos múltiplos utilizado no SAEB. Esse

método foi implementado na PSP em 2007 e melhor organizado pelo próprio Ruben Klein, em 2009.

Entre 2007 e 2012, com um hiato na aplicação da PSP de 2013 a 2016, o *software* proprietário Bilog-MG foi utilizado nas análises por meio do método de calibração separada. Nesse método, testes diferentes com itens em comum eram calibrados individualmente, e os parâmetros de um deles eram então transformados para a escala do teste de referência. A partir de 2017, essa abordagem foi substituída pelo método de grupos múltiplos calibrados simultaneamente, o qual é utilizado tanto pelo Bilog-MG quanto pelo pacote *mirt* com eficiência.

Destarte, com o crescente domínio de análises educométricas, incluindo a TRI, pela equipe da Divisão de Avaliação da Secretaria Municipal de Educação de São Paulo (DA/SME-SP), adotou-se um modelo de análise em paralelo para o processamento estatístico: as análises foram realizadas simultaneamente por duas equipes técnicas. Nesse modelo, cada equipe elabora seu próprio grupo de referência, denominado "Base Clone". O crucial é que os resultados gerados, tanto dos parâmetros dos itens estimados quanto das proficiências, apresentem forte correlação e médias com diferenças mínimas, restritas às casas decimais.

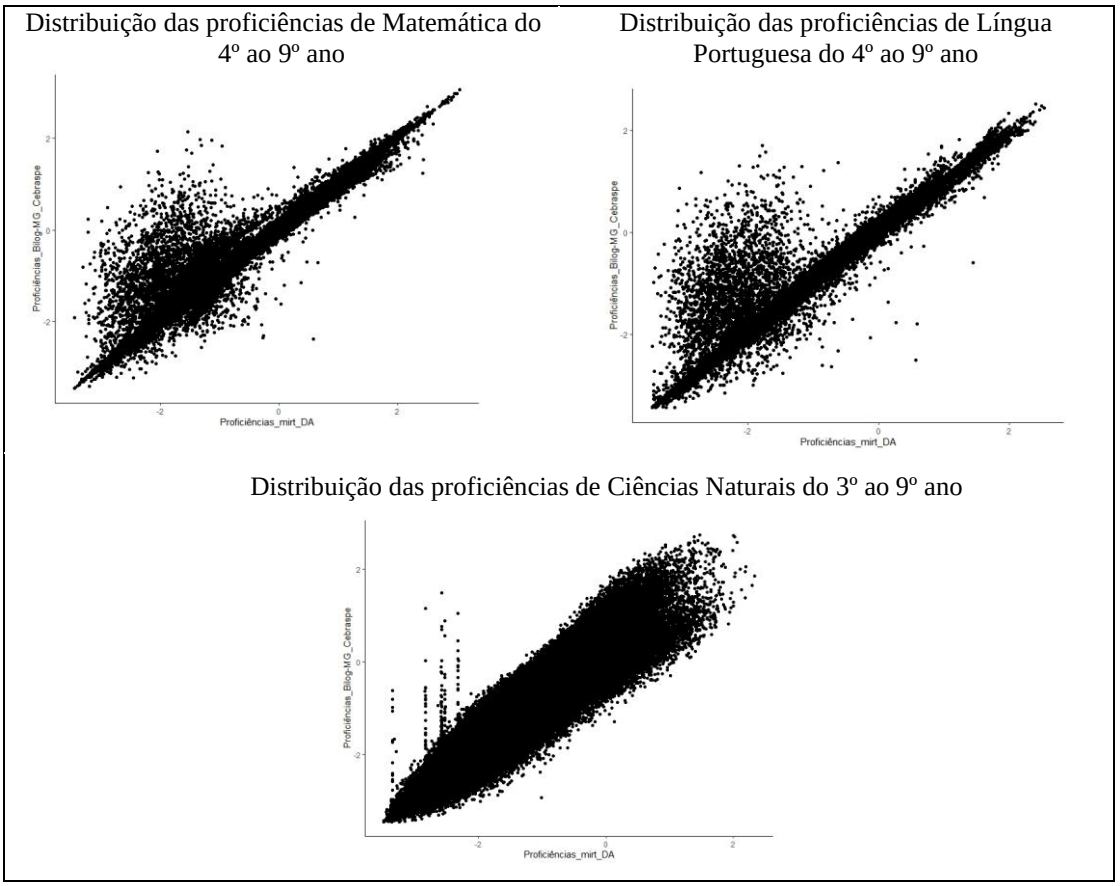
Em 2024, a DA/SME-SP e o Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos (Cebraspe), vencedor da licitação para a PSP 2024, realizaram análises em paralelo. A principal diferença neste ano, além das bases clones distintas para cada parte, residiu na utilização de dois *softwares*/pacotes diferentes para as análises. Assim, o objetivo deste trabalho é apresentar os resultados da comparação entre as medidas obtidas no Bilog-MG e no pacote *mirt.*

3 Resultados

Dado o espaço limitado, apresentam-se aqui alguns aspectos das diversas análises e resultados. A eventual seleção para apresentação em painel permitirá a expansão dessa discussão. Assim, o foco, neste momento, são as comparações entre as proficiências e um exemplo de análise de Funcionamento Diferencial do Item (DIF) entre as duas abordagens, utilizando a Curva Característica do Item (CCI) para verificar a relevância das diferenças. Ressalta-se que as análises do Cebraspe foram atualizadas com a consolidação das bases, o que pode gerar divergências pontuais.

O Quadro 1 apresenta os gráficos das distribuições das proficiências, comparando as duas abordagens de análise para Língua Portuguesa, Matemática e Ciências Naturais. Os três gráficos sintetizam informações relevantes sobre as estimativas de proficiências geradas por ambos os modelos. Observa-se uma maior dispersão na parte inferior da escala, em que a correlação entre as proficiências foi menos intensa. No geral, as correlações entre as distribuições foram elevadas, registrando, respectivamente: 0,984; 0,985 e 0,923.

Quadro 1: **Distribuições das proficiências comparadas entre os modelos analíticos – Prova São Paulo 2024**

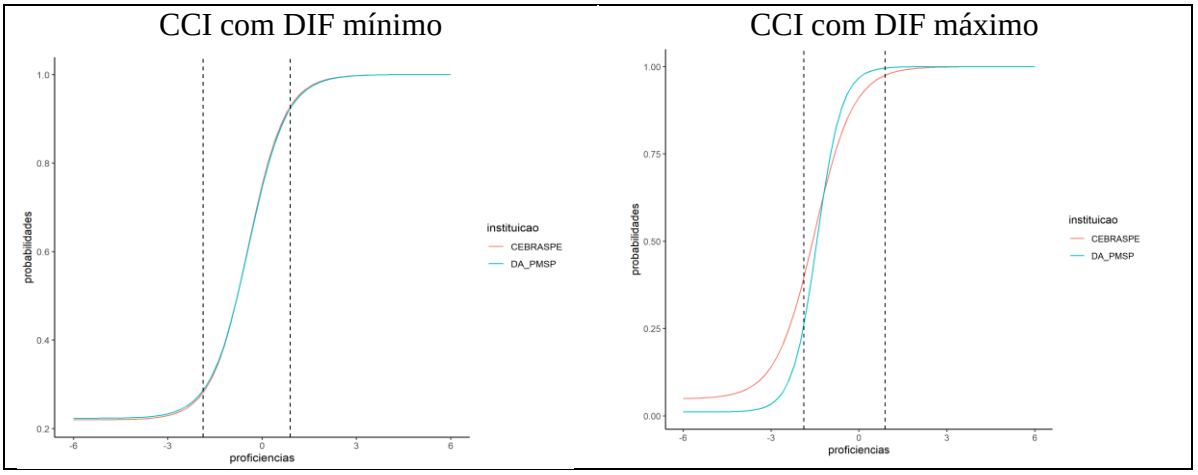


Fonte: SME-DA/Cebraspe.

Conforme proposto em Klein e Ricarte (2025), a comparação entre os dois modelos analíticos não deve se restringir ao cotejamento dos parâmetros dos itens, mas também observar as CCIs. Como exemplo, no Quadro 2, observam-se duas imagens, representando as curvas características de dois itens de Matemática, derivadas dos dois modelos de processamento da PSP 2024. A primeira imagem, referente ao item com menor valor de DIF, demonstra ausência de diferença gráfica significativa entre as probabilidades calculadas para o modelo "CEBRASPE" e para o modelo "DA_PMSP", o que, conforme Klein e Ricarte (2025), sugere a viabilidade da estimativa utilizando o pacote *mirt* em substituição ao Bilog-MG. Contudo, a

segunda imagem, relativa ao item com maior DIF, demonstra uma distância entre as curvas em dois trechos distintos do eixo das proficiências: a maior distância ocorre fora do intervalo entre o 5º e o 95º percentil, enquanto a outra distância significativa se encontra dentro desse intervalo, demandando análise das causas dessa divergência. Apesar disso, a ampla maioria dos itens de Matemática não apresentou DIF, com CCIs semelhantes às da primeira imagem.

Quadro 2: **Curvas Características dos Itens de Matemática: Mínimo e Máximo DIF – Prova São Paulo 2024**



Fonte: SME-DA/Cebraspe.

4 Conclusões e Considerações Finais

De modo geral, a Prova São Paulo de 2024 configura um marco para as análises psicométricas, ou educométricas, ao avançar para o uso de uma linguagem livre por meio de um pacote mundialmente conhecido e com ampla comunidade. Assim, os esforços implementados pela Divisão de Avaliação da SME-SP, como a aplicação de Testes Adaptativos Informatizados (TAI) bimestrais, representam mais um passo na direção de ampliar o uso de ferramentas livres e de compartilhar conhecimento.

5 Referências

CHALMERS, R. P. mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software, v. 48, n. 6, p. 1–29, 2012. Disponível em: https://doi.org/10.18637/jss.v048.i06. Acesso em: 12 maio 2025.

KLEIN, Ruben; RICARTE, T. A. Calibração com parâmetros de itens fixos. Ensaio: Avaliação e Políticas Públicas em Educação, Rio de Janeiro, v. 33, n. 127, e0255144, p. 1-19, abr./jun. 2025.

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.