O Método Angoff na Pré-Calibração de Itens por Nível de Dificuldade: Uma Revisão Baseada em Evidências

Tatiane Gonçalves Moraes
Universidade Federal de Juiz de Fora
Juiz de Fora, MG, Brasil
tatiane.moraes@estudante.ufjf.br

Maria Amélia Cardoso
Poliedro Educação
São José dos Campos, SP, Brasil
Maria.cardoso@sitemapoliedro.com.br

Vanessa de Freitas Santos
Poliedro Educação
São José dos Campos, SP, Brasil
Vanessa.santos@sitemapoliedro.com.br

Resumo

O presente artigo apresenta os resultados da pesquisa conduzida pelo Núcleo de Avaliação do Poliedro Educação sobre a aplicação do método Angoff como técnica de pré-calibração de itens em avaliações em larga escala. A análise está centrada nos dados de Matemática, obtidos a partir da primeira aplicação em uma avaliação inspirada no Pisa. O objetivo principal é demonstrar a viabilidade e a robustez do método Angoff no contexto da educação básica brasileira, contribuindo com evidências para o debate nacional sobre práticas avaliativas inovadoras.

Palavras-chave: Angoff; Pré-teste; Matemática; Teoria de Resposta ao item.

1 Introdução

O método Angoff, proposto por William Angoff (1971), é uma técnica amplamente utilizada para estabelecer padrões de desempenho (cut scores) em avaliações educacionais. Juízes especialistas estimam a probabilidade de um candidato limítrofe acertar cada item. No contexto da Avaliação Bússola, essa técnica foi adaptada para estimar previamente a dificuldade dos itens. Referências como Hambleton e Pitoniak (2006), Plake e Cizek (2012) e Pasquali (2007) fundamentam sua aplicação em contextos de pré-testagem e construção de escalas.

A definição de padrões de desempenho e a ancoragem de itens em escalas de proficiência são aspectos essenciais em avaliações educacionais de larga escala. O método Angoff é amplamente utilizado como referência para estimativas subjetivas de dificuldade e corte, baseando-se na expertise de avaliadores. Tradicionalmente empregado para estabelecer pontos de corte, este método vem sendo ampliado para funções como a pré-calibração de itens, sobretudo quando se deseja garantir um modelo de montagem de teste ancorado em níveis de proficiência antes da aplicação estatística plena.

Este estudo apresenta os fundamentos do método Angoff, discute sua validade e confiabilidade, e descreve a experiência de aplicação desse modelo no processo de pré-testagem dos itens de Matemática, um instrumento desenvolvido pelo Núcleo de Avaliação do Sistema Poliedro de Educação, inspirado no PISA e estruturado com base na Teoria de Resposta ao Item (TRI).

O método Angoff tradicional baseia-se na estimativa, por parte de especialistas, da probabilidade de um candidato limítrofe acertar cada item de um teste. Conforme Livingston e Zieky (1982), esse candidato representa o limiar inferior aceitável de desempenho para determinado nível. O julgamento especializado se dá por meio de rodadas sucessivas, com feedback e discussão, até se atingir consenso. A metodologia tem se mostrado robusta, com coeficientes de confiabilidade interjuízes superiores a 0,80 (Impara & Plake, 1997).

Autores como Hambleton & Pitoniak (2006) e Raymond & Reid (2001) destacam a necessidade de formação e padronização das análises para evitar viés na estimativa. Recentemente, o método tem sido adaptado para auxiliar na ancoragem de itens em escalas de proficiência, como pré-condição à aplicação operacional em larga escala.

2 Metodologia

Inspirada no Programa Internacional de Avaliação de Estudantes (PISA) da OCDE, a Avaliação Bússola foi concebida para avaliar competências além da reprodução de conteúdos,

focando na aplicação do conhecimento em situações reais. Seguindo esse referencial, adaptou-se a metodologia à realidade brasileira, com ênfase no diagnóstico formativo de estudantes da rede particular. A metodologia central é baseada na Teoria de Resposta ao Item (TRI), um modelo matemático amplamente utilizado em avaliações em larga escala, como o PISA, devido à sua capacidade de medir o desempenho dos estudantes de forma precisa e adaptativa (HOGAN, 2006). Neste caso, o modelo da TRI utilizado na Avaliação Bússola, assim como no PISA, foi o de dois parâmetros: dificuldade e discriminação. O parâmetro de dificuldade identifica quão complexo é o item, enquanto o parâmetro de discriminação avalia quão eficaz o item é na diferenciação entre estudantes com níveis variados de habilidade (PASQUALI, 2017). Além disso, para que a experiência para os estudantes fosse a mais proxima do Pisa, utilizou-se a mesma plataforma digital usada por esse programa, garantindo aplicação interativa e segura. Na etapa de pré-testagem de Matemática, empregou-se o método Angoff para estimar a dificuldade dos itens antes da aplicação em larga escala. Cinco especialistas em educação matemática, com experiência em construção de itens e domínio das escalas de proficiência internacionais, analisaram cada questão considerando a complexidade cognitiva (Taxonomia de Bloom), o objeto de conhecimento e o contexto de aplicação

Os juízes estimaram a probabilidade de acerto para um estudante limítrofe em uma escala de 0 a 1, com médias servindo como parâmetro inicial. Itens com divergências foram discutidos em painel até alcançar consenso. Esse processo buscou garantir maior equilíbrio na montagem dos blocos, evitando sobreposição de itens muito fáceis ou muito difíceis, e oferecendo uma matriz inicial coerente com os níveis de proficiência a serem analisados pela TRI na etapa de pós-aplicação.

A pré-calibração por Angoff, nesse contexto, desempenhou um papel estratégico tanto na distribuição equilibrada de dificuldade dos itens quanto na preparação para uma posterior análise psicométrica mais robusta, garantindo que a escala de proficiência fosse representativa, comparável e tecnicamente válida.

A aplicação piloto envolveu mais de 6 mil estudantes da rede particular brasileira, abrangendo todas as regiões do país. Foram avaliadas três áreas centrais: Leitura, Matemática e Ciências e prova contou com 150 itens que foram organizados da seguinte forma:

- Bloco Principal (BPC**)**: Itens de Ciências.
- Bloco Secundário de Leitura (BSL)**:** Itens de Leitura com contextualização científica
- Bloco Secundário de Matemática (BSM)**:** Itens de Matemática com intersecção em Ciências

A distribuição seguiu a metodologia de Blocos Incompletos Balanceados (BIB), com itens de resposta selecionada e construída para avaliar diferentes habilidades. Já a pré-calibração via Angoff assegurou:

- Balanceamento**:** Distribuição equilibrada de dificuldade nos blocos
- Validade**:** Alinhamento com os níveis de proficiência do PISA
- Precisão: Base para a posterior calibração estatística pela TRI

Um sexto especialista validou as estimativas antes do pré-teste, reforçando a confiabilidade do processo. Essa etapa foi crucial para garantir uma escala psicometricamente sólida e comparável a padrões internacionais.

A Avaliação Bússola buscou incorporar os mesmos princípios do Pisa internacional, adaptando-os à realidade brasileira e ao perfil dos estudantes da rede particular de ensino. O objetivo da Bússola é avaliar de maneira mais significativa as competências e habilidades adquiridas por estudantes ao longo da escolarização, oferecendo uma leitura diagnóstica e formativa sobre suas trajetórias de aprendizagem, suas potencialidades e lacunas.

3 Resultados

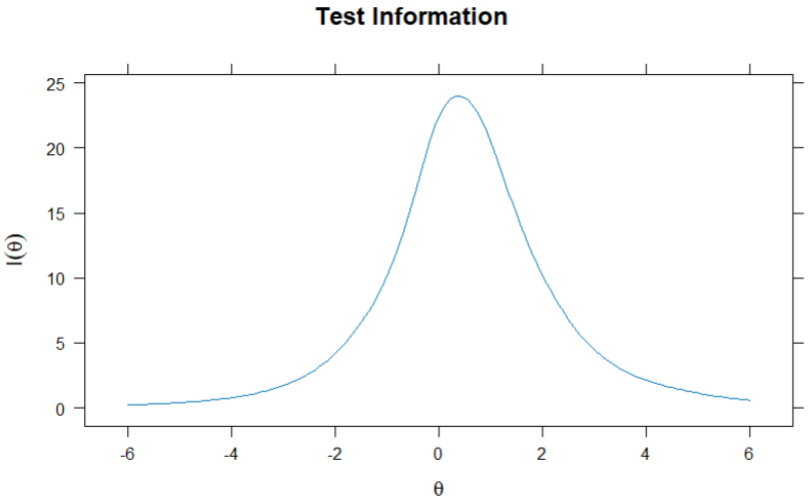
A prova de Matemática da Bússola foi aplicada a uma amostra de mais de 6 mil estudantes. O modelo adotado foi a TRI de dois parâmetros (discriminação e dificuldade). Os resultados, nesta área, indicaram uma média de proficiência de 456 pontos (DP = 100), alinhada ao desempenho da rede privada no PISA 2022. Os dados psicométricos revelaram:

- Correlação item-teste média: 0,44;
- Itens com bisseriais abaixo de 0,2 foram considerados para exclusão ou revisão;
- Um total de 4 itens apresentou comportamento anômalo (CCI invertida), o que indica erro de gabarito ou problema de formulação;
- Distribuição de proficiência concentrada entre os níveis 3 e 5.

Os resultados da análise estatística revelaram que a maioria dos itens apresentou bons índices de discriminação (acima de 0,4) e dificuldade variada, compondo uma escala adequada. Contudo, alguns itens apresentaram comportamento anômalo na CCI, como aumento da probabilidade de erro com maior proficiência. Esses itens foram identificados para revisão ou substituição na reaplicação prevista para 2025. A aplicação do método Angoff mostrou-se eficaz na pré-calibração, pois houve boa correlação entre as estimativas dos especialistas e os parâmetros empíricos.

Observa-se que o teste apresentou uma dificuldade de mediana ou um pouco acima da média, visto que a curva de informação do teste está centrada um pouco à direita da média de

proficiência original (0) e apresenta valores mais elevados entre -2 e 2 da escala dada. Isso demonstra que a distribuição de itens foi eficaz em cobrir amplamente o espectro de desempenho dos alunos avaliados.



Esses resultados reforçam a robustez metodológica do processo de pré-calibração conduzido pelo Núcleo de Avaliação do Poliedro. A boa correspondência entre os julgamentos prévios dos especialistas, ancorados pelo método Angoff, e os parâmetros psicométricos estimados após a aplicação, atesta a validade do processo, especialmente em contextos em que a testagem empírica ainda não foi consolidada. Além disso, a utilização da TRI com dois parâmetros possibilitou a identificação criteriosa de itens com baixa discriminação ou comportamento inconsistente, permitindo ajustes estratégicos no banco de itens antes da reaplicação. Em suma, a articulação entre a ancoragem técnica prévia via método Angoff e a análise estatística por meio da TRI consolidou a Avaliação Bússola como uma experiência inovadora e tecnicamente consistente no cenário nacional. A estrutura do teste revelou-se sensível à diversidade de perfis dos estudantes, e ao mesmo tempo rigorosa no controle da qualidade psicométrica, aspectos essenciais para qualquer avaliação de larga escala que deseje aliar inovação, comparabilidade e responsabilidade técnica.

4 Conclusões e Considerações Finais

A utilização do método Angoff como etapa de pré-calibração de itens demonstrou ser válida e confiável no contexto da Avaliação Bússola. Os resultados reforçam a importância de associar expertise técnica e estatística no processo de construção e validação de instrumentos avaliativos. O cruzamento entre as estimativas subjetivas e os dados empíricos forneceu subsídios para aprimoramento do banco de itens e maior precisão na mensuração da proficiência dos estudantes. Além disso, a análise dos dados da área de Matemática

evidenciou que os itens previamente ancorados apresentaram, em sua maioria, bons índices de discriminação e dificuldade, compondo uma escala coerente com o perfil da população-alvo. A presença de poucos itens com comportamento anômalo, rapidamente identificados pela análise da Curva Característica do Item (CCI) e dos coeficientes bisseriais, demonstra a eficácia do processo de pré-teste conduzido.

A boa correlação entre os julgamentos realizados pelos especialistas e os parâmetros estimados estatisticamente após a aplicação reforça o potencial do método Angoff como ferramenta estratégica de pré-calibração, especialmente em avaliações com inspiração internacional e que exigem comparabilidade e rigor técnico, como é o caso da Bússola.

Por fim, o uso articulado da TRI com o método Angoff consolidou uma metodologia sólida de desenvolvimento de itens e montagem de testes. Tal abordagem promove uma avaliação mais justa, sensível às diferenças de desempenho entre os estudantes, e oferece ao campo da avaliação educacional no Brasil um exemplo promissor de inovação com responsabilidade técnica e psicométrica. O processo agora avança para a fase de reaplicação, momento em que os itens serão refinados e novas evidências serão geradas para ampliar a robustez e a validade do instrumento.

5 Referências

ANGOFF, William H. *Scales, norms and equivalent scores*. Princeton, NJ: Educational Testing Service, 1971.

HAMBLETON, Ronald K.; PITONIAK, Mary J. Setting performance standards. In: BRENT, Robert L. *Educational Measurement*. 4. ed. Westport, CT: American Council on Education, 2006. p. 433–470.

HOGAN, Thomas P. *Introdução à prática de testes psicológicos*. Rio de Janeiro: LTC, 2006.

IMPARA, James C.; PLAKE, Barbara S. Standard setting: an alternative approach. *Journal of Educational Measurement*, v. 34, n. 4, p. 353–366, 1997.

LIVINGSTON, Samuel A.; ZIEKY, Michael J. *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton: Educational Testing Service, 1982.

PASQUALI, Luiz. *Psicometria: teoria dos testes na psicologia e na educação*. Petrópolis: Vozes, 2007.

PLAKE, Barbara S.; CIZEK, Gregory J. *Setting performance standards: Foundations, methods, and innovations*. New York: Routledge, 2012.

RAYMOND, Mark R.; REID, Jeanne B. Who made thee a judge? Selecting and training participants for standard setting. *CSE Technical Report 539*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), 2001.