Instrumentos padronizados para avaliação dos conhecimentos de Matemática de ingressantes no ensino superior nas áreas de Ciências Exatas e Engenharias

Henrique Antonio Mendonça Faria [1]
Jorge Manuel Vieira Capela[2]
Marisa Veiga Capela[3]
Fabio Roberto Chavarette[4]

RESUMO

Ao ingressar nos cursos de ciências exatas e engenharias do ensino superior, a maioria dos estudantes enfrenta dificuldades consideráveis em Matemática. Esse problema tem-se agravado nos últimos anos, acarretando prejuízos sociais para os alunos e institucionais para as universidades. Ações de reforço educacional para os ingressantes são comuns em diversas instituições. Contudo, para garantir a eficácia dessas medidas, é relevante identificar com maior precisão as lacunas nos conteúdos matemáticos para cada grupo universitário local. Este estudo propõe conceituar e identifica instrumentos avaliativos padronizados, aplicáveis no primeiro ano do ensino superior, com o objetivo de identificar as deficiências dos estudantes em conteúdos da Matemática. A principal metodologia adotada foi a pesquisa bibliográfica exploratória. A revisão da literatura revelou que instrumentos padronizados de avaliação são utilizados desde o início do século XX e continuam sendo aperfeiçoados até os dias atuais. Os instrumentos padronizados avaliados neste estudo baseiam-se em pressupostos da Psicometria e na Teoria da Resposta do Item (TRI). O modelo de George Rasch (1911-1980) é usado para estimar o nível de conhecimento, competência e aptidão dos indivíduos em determinados conteúdos de Matemática, utilizando as características dos itens avaliativos. Este modelo possibilita estimativas quantitativas, tanto dos parâmetros discriminativos do domínio dos conteúdos, quanto das possibilidades de acerto ao acaso de itens com respostas em múltipla escolha. A TRI e o modelo de Rasch são apontados pela literatura como os mais adequados para subsidiar a construção de instrumentos avaliativos em populações heterogêneas. O instrumento elaborado e avaliado segundo esses pressupostos poderá auxiliar os professores de matemática e as instituições de ensino, na avaliação, identificação de dificuldades e proposição de soluções de reforço no contexto da sala de aula. Assim, as soluções poderão ser mais assertivas para aprimorar o processo de ensino-aprendizagem nos anos iniciais dos cursos superiores em ciências exatas.

Palavras-chave: Ensino Superior, Conhecimento em Matemática, Teste padronizado.

[1] Professor orientador: Doutor em Ciências, IQ Unesp - SP, henrique.faria@unesp.br.
[2] Professor orientador: Doutor em Biotecnologia, IQ Unesp - SP, jorge.capela@unesp.br.
[3] Professora orientadora: Doutora em Química, IQ Unesp - SP, marisa.capela@unesp.br.
[4] Professora orientador: Doutor em Engenharia, IQ Unesp - SP, fabio.chavarette@unesp.br.

INTRODUÇÃO

Os testes padronizados são instrumentos de avaliação utilizados para medir habilidades, conhecimentos, aptidões e outros atributos de indivíduos de maneira uniforme e consistente. Eles são projetados para garantir que todos os examinados sejam avaliados em condições semelhantes, utilizando os mesmos critérios de correção e interpretação. Esses testes são amplamente utilizados em contextos educacionais, profissionais e psicológicos, oferecendo uma maneira objetiva de comparar o desempenho entre diferentes indivíduos ou grupos. A elaboração, aplicação e avaliação dos resultados de um teste estão inseridos em uma grande área da psicologia chamada de psicometria (Gregory, 2010). Os testes desempenham um papel crucial na avaliação e compreensão das capacidades cognitivas, comportamentais e emocionais dos indivíduos. Eles são desenvolvidos com rigor científico para garantir que suas aplicações sejam válidas, confiáveis e justas para toda a população avaliada.

Na educação esses testes oferecerem uma medida uniforme para avaliar o desempenho acadêmico dos estudantes. Eles possibilitam a comparação de resultados entre diferentes grupos de estudantes, escolas e até sistemas educacionais, promovendo a equidade e a justiça na avaliação educacional (Urbina, 2004). Esses testes são amplamente utilizados para diversas finalidades, incluindo (a) **Avaliação de aprendizagem**: ajudam a determinar o quanto os estudantes aprenderam em relação ao currículo estabelecido, identificando áreas de força e fraqueza; (b) **Decisões educacionais**: os resultados desses testes são frequentemente usados para tomar decisões importantes, como promoção de série, atribuição de recursos e desenvolvimento de programas de melhoria escolar; (c) **Responsabilização**: permitem que escolas e professores sejam responsabilizados pelo desempenho dos estudantes, incentivando a melhoria contínua da qualidade do ensino. (d) **Admissões e certificações**: são essenciais em muitos processos de admissão em universidades e programas de certificação profissional.

Uma das principais vantagens dos testes padronizados é a objetividade. Eles são aplicados e pontuados de maneira uniforme, minimizando a influência de vieses subjetivos. Isso garante que os resultados sejam consistentes independentemente de quem administra o teste. Outra característica é que os testes padronizados permitem comparar os resultados de diferentes indivíduos ou grupos em diferentes momentos e locais. Isso é essencial para pesquisas longitudinais e estudos comparativos entre

populações diversas. No contexto educacional, os testes padronizados ajudam a identificar as necessidades individuais dos alunos, contribuindo para o desenvolvimento de estratégias pedagógicas personalizadas. Eles também são utilizados para orientar decisões sobre carreiras e desenvolvimento pessoal.

A história dos testes psicológicos padronizados remonta ao final do século XIX e início do século XX, marcando uma evolução significativa nas metodologias de avaliação. Os marcos históricos dessa evolução foram apresentados por Gregory (2010). Francis Galton (1822–1911) foi o pioneiro nos estudos de diferenças individuais, ele foi um dos primeiros a desenvolver uma bateria de testes que mediam habilidades sensoriais e motoras. Seu trabalho lançou as bases para a quantificação das capacidades humanas. Reconhecido por fundar o primeiro laboratório de psicologia em 1879, Wilhelm Wundt (1832–1920) utilizou métodos experimentais para estudar processos mentais, enfatizando a medição objetiva e a reprodutibilidade dos experimentos.

Influenciado por Galton, James McKeen Cattell (1860–1944) introduziu o conceito de "testes mentais" e destacou a importância da experimentação e medição na psicologia. Seu trabalho estabeleceu a agenda moderna para o desenvolvimento de testes psicológicos. Nessa linha de pesquisa, os testes de inteligência desenvolvidos por Alfred Binet e Théodore Simon em 1905, foram os primeiros a avaliar a inteligência de forma sistemática. No campo educacional, E. L. Thorndike e Joseph Rice tiveram contribuições significativas ao desenvolverem métodos para medir a eficiência da aprendizagem nas escolas, criando escalas que podiam comparar o desempenho dos estudantes em habilidades específicas. Em 1910, Thorndike publicou sua escala de escrita, que foi seguida por testes padronizados para avaliar habilidades em aritmética, leitura e ortografia.

A revisão de Lewis Terman em 1916 resultou no Stanford-Binet, amplamente utilizado para medir a inteligência nos Estados Unidos. Na década de 1920, exames objetivos baseados no *Army Alpha*, um teste militar, começaram a ser usados juntamente com as notas do ensino médio para decisões de admissão em faculdades e universidades, culminando na criação do *Scholastic Aptitude Test* (SAT)

METODOLOGIA

A presente pesquisa teve como principal objetivo analisar a literatura científica sobre a utilização de testes padronizados de habilidades matemáticas em processos seletivos, com foco em seus impactos na equidade e na eficácia desses processos. Para alcançar esse objetivo, foi realizada uma revisão sistemática da literatura, seguindo os seguintes procedimentos metodológicos. Inicialmente, foi definida uma pergunta de pesquisa clara e precisa: qual o impacto da utilização de testes padronizados de habilidades matemáticas em processos seletivos, considerando aspectos como equidade e eficácia? Essa pergunta norteou todo o processo de busca e seleção dos estudos.

Em seguida, foi realizada uma busca em bases de dados acadêmicas como Scielo, *Web of Science*, e Google Scholar. Foram utilizados descritores específicos, como "testes padronizados", "habilidades matemáticas", "processos seletivos", "equidade" e "eficácia", combinados com operadores booleanos para refinar a busca. A seleção dos estudos incluídos na revisão foi realizada com base em critérios relevância. Foram incluídos estudos originais, publicados em periódicos indexados, que investigaram a relação entre o desempenho em testes padronizados de habilidades matemáticas, considerando aspectos como a equidade e a eficácia desses processos.

A análise dos dados coletados foi realizada de forma qualitativa, utilizando a técnica de análise temática. É importante ressaltar que esta pesquisa não envolveu a coleta de dados primários com seres humanos. Portanto, não houve necessidade de aprovação por um comitê de ética.

REFERENCIAL TEÓRICO

Testes para Habilidades Matemáticas e de Raciocínio Lógico

Os testes de habilidades matemáticas e de raciocínio lógico são fundamentais para avaliar a capacidade dos indivíduos em resolver problemas quantitativos e aplicar o pensamento lógico em diversas situações. Esses testes são utilizados em contextos educacionais e profissionais para medir o nível de proficiência em matemática e o potencial dos estudantes ou candidatos. Os testes focam na avaliação de conhecimentos e competências em áreas como aritmética, álgebra, geometria, estatística e cálculo. Os testes podem ainda ser subdivididos em: teste de raciocínio quantitativo; testes de progresso em matemática; testes de raciocínio lógico; testes de raciocínio abstrato e testes de raciocínio dedutivo e indutivo.

Segundo Urbina (2004) os testes de raciocínio quantitativo, componente presentes no *Graduate Record Examination* (GRE) e no *Graduate Management Admission Test* (GMAT) americanos, medem a capacidade do indivíduo em interpretar e analisar dados quantitativos, resolver problemas matemáticos complexos e aplicar conceitos matemáticos em situações práticas. Eles avaliam habilidades em áreas como álgebra, geometria, aritmética e análise de dados.

Os testes de progresso em matemática são utilizados para monitorar o desenvolvimento das habilidades matemáticas dos estudantes ao longo do tempo. Esses testes são aplicados em diferentes níveis educacionais para avaliar o domínio de conceitos matemáticos fundamentais e identificar áreas que precisam de melhorias.

Os testes de raciocínio lógico são projetados para avaliar a capacidade dos indivíduos de pensar de maneira lógica e estruturada, identificar padrões, resolver problemas e fazer inferências lógicas. Estes testes são utilizados em processos de seleção acadêmica e recrutamento profissional.

Os testes de raciocínio abstrato medem a capacidade de entender conceitos não verbais e resolver problemas usando o pensamento lógico. Esses testes são frequentemente utilizados para avaliar a aptidão dos candidatos para cargos que exigem fortes habilidades analíticas e de resolução de problemas. Testes como o *Raven's Progressive Matrices* americano são exemplos clássicos de testes de raciocínio abstrato, utilizados em contextos educacionais e ocupacionais.

Os testes de raciocínio dedutivo avaliam a capacidade de aplicar regras gerais a situações específicas para chegar a conclusões lógicas, enquanto os testes de raciocínio indutivo medem a capacidade de identificar padrões e fazer generalizações a partir de dados específicos.

No contexto corporativo, testes de raciocínio lógico são frequentemente utilizados como parte do processo de seleção para identificar candidatos com habilidades analíticas fortes. Esses testes ajudam as empresas a selecionar indivíduos que são capazes de resolver problemas complexos e tomar decisões baseadas em lógica e análise. Empresas frequentemente utilizam testes de raciocínio lógico como parte de suas avaliações de aptidão para garantir que os candidatos possuam as habilidades necessárias para desempenhar funções que exigem pensamento crítico e resolução de problemas Urbina (2004).

Testes de múltipla escolha

Os testes de múltipla escolha (TMEs) são instrumentos de avaliação amplamente utilizados em diversos contextos educacionais e profissionais. Eles consistem em perguntas ou itens que oferecem ao examinado várias opções de respostas, das quais apenas uma é correta (Haladyna e Rodriguez, 2013). Esses testes são projetados para medir o conhecimento ou habilidades em uma área específica, oferecendo uma forma padronizada e objetiva de avaliação. Devido à sua estrutura, os TMEs são especialmente eficazes para avaliar conhecimentos factuais, compreensão de conceitos e habilidades de análise e de aplicação.

Os TMEs são caracterizados por sua objetividade, uma vez que a resposta correta é previamente determinada, eliminando a subjetividade no processo de correção. Isso permite uma correção rápida e eficiente, muitas vezes realizada automaticamente por meio de softwares. Além disso, a padronização dos itens facilita a comparação de resultados entre diferentes grupos de examinados, garantindo equidade na avaliação. Além disso, ao elaborar um TME, é crucial evitar pistas não intencionais que possam ajudar os alunos a identificar a resposta correta sem o conhecimento necessário. Isso inclui evitar padrões de resposta previsíveis, linguagem inconsistente entre as alternativas e o uso de expressões absolutas, como "sempre" ou "nunca", que podem sugerir a resposta correta.

Apesar de sua popularidade e utilidade, os testes de múltipla escolha apresentam algumas limitações, segundo Haladyna e Rodriguez (2013). Uma das críticas mais comuns é que esses testes tendem a favorecer a memorização em detrimento de uma compreensão profunda dos conceitos. Como as respostas são pré-definidas, os alunos podem ser induzidos a se concentrar em lembrar fatos e detalhes específicos em vez de desenvolver uma compreensão mais global e crítica do conteúdo.

Outra limitação é a dificuldade de elaborar itens que avaliem efetivamente habilidades cognitivas superiores. Embora seja possível criar TMEs que abordem análise, síntese e avaliação, isso exige um nível elevado de habilidade na redação de questões e uma compreensão profunda do conteúdo a ser avaliado. Além disso, a elaboração de bons distratores é um desafio constante, pois estes devem ser plausíveis o suficiente para que alunos bem preparados possam considerá-los como possíveis respostas corretas.

Os TMEs também podem ser vulneráveis a vieses, tanto culturais quanto contextuais. Um item mal formulado pode favorecer ou prejudicar certos grupos de examinados, dependendo de sua experiência cultural, socioeconômica ou educacional. Isso pode resultar em avaliações que não refletem com precisão as habilidades ou o conhecimento do indivíduo, mas sim sua familiaridade com o formato do teste ou com o conteúdo específico do item (Haladyna e Rodriguez, 2013). Grande parte dessas desvantagens podem ser evitadas na elaboração cuidadosa do teste.

Apesar de algumas limitações, os testes de múltipla escolha oferecem várias vantagens frente aos exames dissertativos tornando-os uma escolha adequada em avaliações de larga escala. Uma das principais vantagens é a objetividade na correção, que garante que todos os examinados sejam julgados com base nos mesmos critérios. Isso é especialmente importante em exames de alta relevância, como vestibulares ou provas de certificação profissional, onde a equidade é fundamental.

Outra vantagem significativa dos TMEs é sua abrangência. Eles permitem a cobertura de uma ampla gama de conteúdo em um único exame, proporcionando uma visão ampla do conhecimento do estudante em diversas áreas. Além disso, a correção automatizada economiza tempo e recursos, permitindo que os resultados sejam disponibilizados rapidamente.

Os TMEs também são flexíveis, podendo ser adaptados para avaliar uma variedade de níveis de dificuldade e tipos de conhecimento. Eles podem ser usados em testes de diagnóstico, avaliações formativas e somativas, e em contextos de ensino e aprendizagem adaptativos, onde o *feedback* imediato é essencial para o progresso do aluno.

A elaboração de um teste de múltipla escolha é um processo que exige cuidado e atenção para garantir a validade e a confiabilidade do instrumento. Monteiro (2010) destaca que a construção de um TME deve começar com a definição clara dos objetivos de aprendizagem que o teste pretende avaliar. A partir desses objetivos, os itens são criados de maneira a refletir as habilidades e conhecimentos que se deseja medir.

Na fase de construção dos itens, é importante garantir que as perguntas sejam formuladas de maneira clara e precisa, evitando ambiguidades que possam confundir os avaliando. Além disso, as alternativas de resposta devem ser cuidadosamente elaboradas. As alternativas incorretas, conhecidas como distratores, desempenham um papel crucial, pois devem ser plausíveis para evitar que os alunos possam facilmente eliminá-las e acertar a resposta correta por exclusão. Os distratores eficazes são aqueles

que representam erros comuns ou concepções equivocadas sobre o tema, desafiando o estudante a demonstrar verdadeiro entendimento do conteúdo.

Outro aspecto importante na elaboração de TMEs é a revisão e validação dos itens antes de sua aplicação. Isso pode envolver a análise dos itens por especialistas na área de conteúdo e a realização de testes-piloto para identificar possíveis falhas ou ambiguidades. A revisão também pode ajudar a garantir que os itens estejam livres de vieses culturais ou linguísticos que possam prejudicar certos grupos de alunos.

Além de medir o conhecimento conceitual, os TMEs podem ser usados para avaliar o raciocínio lógico e as habilidades de resolução de problemas. Questões que envolvem cenários hipotéticos ou que requerem que os alunos façam inferências a partir de dados fornecidos podem ajudar a medir a capacidade dos alunos de aplicar seu conhecimento em contextos novos e desconhecidos. No entanto, a capacidade de um TME de avaliar tais habilidades depende em grande parte da clareza e precisão com que as perguntas são formuladas, bem como da relevância das alternativas de resposta Monteiro (2010). A inclusão de perguntas que exigem análise crítica ou julgamento pode aumentar a validade do teste como uma medida do pensamento complexo.

Os testes de múltipla escolha são ferramentas poderosas e versáteis na avaliação educacional, capazes de medir uma variedade de habilidades e conhecimentos de forma objetiva e eficiente. No entanto, para maximizar sua eficácia, é essencial que sejam cuidadosamente elaborados, com atenção à clareza das perguntas, à plausibilidade dos distratores e à relevância do conteúdo avaliado. Apesar das críticas e limitações, como a ênfase na memorização e a dificuldade em medir habilidades de análise e aplicação, os TMEs continuam a ser uma das formas mais utilizadas de avaliação, devido às suas numerosas vantagens em termos de objetividade, eficiência e flexibilidade.

Testes consolidados em larga escala

O interesse pelas práticas matemáticas dos alunos tem crescido ao longo dos últimos 40 anos Monteiro (2010). Estes estudos focaram nas habilidades cognitivas e estratégias usadas por especialistas em matemática e adultos em contextos informais, ou seja, fora do ambiente escolar. A partir desta linha de pesquisa, emergiu uma compreensão das práticas matemáticas, delineando os comportamentos específicos envolvidos no raciocínio matemático e na resolução de problemas. Dois instrumentos relevantes no contexto educacional atual, o *National Assessment of Educational*

Progress (NAEP) e o Exame Nacional do Ensino Médio (ENEM), apresentam características fundamentais consolidadas para a elaboração de testes.

NAEP

O *National Assessment of Educational Progress* (NAEP) é um instrumento norte-americano utilizado para avaliar o desempenho acadêmico dos estudantes em diversas áreas, incluindo matemática (NAEP, 2021). Desde 1973, o NAEP oferece um panorama nacional do conhecimento e habilidades dos alunos americanos nessa disciplina. Com o objetivo de acompanhar as evoluções na educação, o NAEP atualiza periodicamente seus parâmetros para avaliar o que os estudantes devem saber e ser capazes de fazer em matemática. O foco está em garantir que a avaliação reflita as mudanças na pesquisa, políticas educacionais e práticas pedagógicas.

Para desenvolver os itens da avaliação, o NAEP envolve especialistas em educação matemática, pesquisadores, professores e outros profissionais da área. O propósito é criar um instrumento acessível a todos os estudantes, independentemente de sua origem socioeconômica, étnica, cultural ou linguística. O exame busca compreender não apenas o conteúdo matemático adquirido pelos alunos, mas também sua capacidade de aplicar esse conhecimento em diferentes contextos. Os resultados do NAEP são utilizados por educadores, políticos, pais e o público em geral para compreender o nível de proficiência dos alunos em matemática (NAEP, 2021). Além disso, esses dados servem como base para o desenvolvimento de estratégias de melhoria na educação.

O conteúdo matemático avaliado pelo NAEP abrange cinco grandes áreas: Propriedades e Operações Numéricas; Medidas; Geometria; Análise de dados, Estatística e Probabilidade; e Álgebra. Embora essas áreas sejam distintas, é importante ressaltar que frequentemente há interseções entre elas. Por exemplo, cálculos aritméticos são essenciais em diversas áreas, como Medida e Geometria.

O NAEP enfatiza a importância do pensamento numérico, incluindo compreensão de diferentes representações numéricas, cálculo mental e estimado, e resolução de problemas envolvendo proporções e porcentagens. A avaliação também aborda a medida, com foco em atributos como comprimento, área, volume, tempo e temperatura. Além disso, o exame inclui questões sobre Geometria, envolvendo propriedades de figuras planas e espaciais, bem como transformações geométricas.

A área de Análise de Dados, Estatística e Probabilidade é crucial para desenvolver habilidades de coleta, organização, análise e interpretação de dados. Os

alunos são avaliados em sua capacidade de criar representações gráficas de dados, calcular medidas estatísticas e compreender conceitos probabilísticos. Por fim, a Álgebra é abordada com questões sobre expressões algébricas, equações, funções e relações.

É importante destacar que o NAEP busca avaliar a aplicação do conhecimento matemático em situações cotidianas, além da compreensão de conceitos teóricos. A avaliação inclui itens que exigem dos estudantes a capacidade de modelar situações reais utilizando ferramentas matemáticas, interpretar resultados e comunicar suas ideias matematicamente. Em suma, o NAEP oferece uma visão abrangente do conhecimento matemático dos estudantes, considerando tanto o conteúdo específico quanto a aplicação prática desse conhecimento. Ao avaliar essas diferentes dimensões, o exame contribui para a compreensão do desempenho matemático dos alunos e fornece informações valiosas para a melhoria da educação matemática.

A pesquisa em educação matemática também experimentou uma "virada social" (Forzani et al., 2022), marcada por um movimento em direção ao estudo da aprendizagem matemática inserida em atividades sociais, incluindo práticas discursivas. Isso implica que os estudantes utilizam seu conhecimento e habilidades matemáticas em diversos contextos sociais, como na escola, em casa ou em jogos com amigos. O quadro de referência de matemática do NAEP 2026 reflete essa visão mais abrangente e completa do que significa conhecer e praticar matemática. Pela primeira vez, as práticas matemáticas foram incorporadas como um componente fundamental da avaliação do NAEP. Estas práticas matemáticas do NAEP foram selecionadas após uma revisão da literatura atual, de quadros de referência de avaliações nacionais e internacionais, bem como de padrões nacionais e estaduais norte americanos. Destacam-se cinco práticas principais: representação; abstração e generalização; justificação e prova; modelagem matemática e colaboração em matemática.

A *representação* envolve o reconhecimento, uso, criação, interpretação ou tradução entre representações adequadas ao nível de escolaridade e ao conteúdo matemático avaliado. Esta prática enfatiza a habilidade dos estudantes em abordar diferentes formas de representação matemática, como gráficos, tabelas, equações e modelos visuais, que são cruciais para a compreensão e comunicação de ideias matemáticas.

A *abstração e generalização* refere-se à capacidade de contextualizar, identificar semelhanças entre casos, itens ou problemas e estender o raciocínio para um domínio

mais amplo, de acordo com o nível de escolaridade e o conteúdo matemático avaliado. Esta prática é central para o desenvolvimento de habilidades de pensamento abstrato, permitindo que os alunos identifiquem padrões e generalizem conceitos matemáticos para novas situações.

Justificação e prova inclui a criação, avaliação, apresentação ou refutação de argumentos matemáticos, com maneiras apropriadas ao desenvolvimento e ao conteúdo matemático. Justificar e provar são componentes essenciais do pensamento matemático rigoroso, exigindo que os estudantes sustentem suas conclusões com argumentos lógicos e baseados em evidências.

Na *modelagem matemática* estão presentes a compreensão de um cenário e a identificação de um problema a ser resolvido. Em seguida, a matematizá-lo permite encontrar uma solução adequada para o entendimento do fenômeno ou processo. Por fim é necessária a verificação e a viabilidade do modelo. Essa prática enfatiza a aplicação da matemática em contextos do mundo real, permitindo que os estudantes utilizem ferramentas matemáticas para resolver problemas complexos e interpretar os resultados em termos de sua aplicabilidade.

O quinto item, a *colaboração em matemática* diz respeito ao empreendimento social de fazer matemática em conjunto com outros, por meio de discussões e resolução colaborativa de problemas, onde ideias são apresentadas, debatidas, conectadas e desenvolvidas em direção a uma solução e compreensão compartilhada. A colaboração em matemática envolve o pensamento conjunto entre indivíduos para construir uma solução para o problema de forma apropriada ao desenvolvimento.

As práticas matemáticas contribuem para o raciocínio matemático, mas também dependem de outras habilidades, como o domínio conceitual e a fluência computacional. Da mesma forma, as práticas matemáticas podem contribuir para a compreensão conceitual, mas não são intercambiáveis com ela (Forzani et al., 2022). A compreensão conceitual refere-se ao conhecimento das estruturas e relações subjacentes na matemática que transcendem a aplicação de algoritmos familiares, enquanto as práticas são fluidas e responsivas tanto a problemas familiares quanto desconhecidos.

O aumento da ênfase nas práticas matemáticas é evidente nos padrões estaduais e nacionais dos testes norte americanos. Atualmente, é amplamente aceito que saber e fazer matemática envolve práticas como generalizar, conjecturar, justificar, matematizar, resolver problemas, comunicar e dar sentido aos conceitos. À medida que os estudantes lidam e discutem ideias e problemas matemáticos, individualmente e em

grupos, eles se envolvem nessas práticas, que os familiarizam com as normas de fazer matemática. A inclusão das práticas matemáticas do NAEP não está separada do conteúdo matemático, ela indica mudança significativa no quadro de referência de matemática do NAEP, com o objetivo de garantir que a avaliação capture uma imagem completa e abrangente do conhecimento e das habilidades matemáticas dos estudantes.

ENEM

O Exame Nacional do Ensino Médio (ENEM) foi criado em 1998 no Brasil com o objetivo de avaliar o desempenho dos estudantes ao final da escolaridade básica, medindo o desenvolvimento das competências e habilidades necessárias ao exercício pleno da cidadania (Andriola, 2011). Inicialmente, o exame contava com um número reduzido de participantes, mas sua importância e abrangência cresceram significativamente ao longo dos anos. Em 2008, o ENEM atingiu mais de quatro milhões de inscritos, e em 2010, esse número ultrapassou 4,6 milhões. Além de avaliar o desempenho dos estudantes, o ENEM passou a ser utilizado como critério de seleção para o Programa Universidade para Todos (ProUni) a partir de 2004 e, posteriormente, por diversas Instituições de Ensino Superior (IES) como parte de seus processos seletivos.

Em 2010, o Ministério da Educação (MEC) propôs uma reformulação do ENEM, tornando-o um instrumento de seleção unificada para as IES. O novo ENEM passou a ser composto por provas que avaliam quatro áreas do conhecimento: linguagens, códigos e suas tecnologias (incluindo redação); ciências humanas e suas tecnologias; ciências da natureza e suas tecnologias; e matemática e suas tecnologias. Cada área é composta por 45 questões de múltipla escolha, aplicadas em dois dias. A redação deve ser em língua portuguesa, em formato dissertativo-argumentativo, abordando temas de ordem social, científica, cultural ou política.

Uma das principais diferenças entre o antigo e o novo ENEM é a estrutura das provas. Até 2008, o exame era composto por 63 itens interdisciplinares, sem articulação direta com os conteúdos do ensino médio e sem a possibilidade de comparar os resultados dos estudantes ao longo do tempo. O novo formato permite essa comparação, o que possibilita a organização de séries históricas que podem contribuir para análises educacionais. Além disso, a nova versão do exame busca não apenas avaliar o conhecimento dos estudantes, mas também suas habilidades e competências em enfrentar situações-problema, construir argumentações e elaborar propostas.

Em 2015, houve uma ampliação no banco de itens, buscando maior diversidade nas questões e no estilo das perguntas. A estrutura básica do exame, no entanto, manteve-se com as 180 questões e a redação. A partir de 2025, devido à adoção do Novo Ensino Médio no contexto brasileiro, o ENEM poderá passar por mudanças significativas.

As IES possuem autonomia para decidir como utilizar o ENEM em seus processos seletivos, podendo optar por usá-lo como fase única, em combinação com o vestibular, ou como fase única para vagas remanescentes. O Sistema de Seleção Unificada (Sisu) foi criado para gerenciar esse processo de seleção, permitindo que os candidatos sejam selecionados exclusivamente com base nas notas obtidas no ENEM. O exame, portanto, tornou-se um modelo de avaliação adaptado às necessidades de uma sociedade em constante transformação, focando no desenvolvimento de competências essenciais para o cidadão do século XXI.

O ENEM busca avaliar competências complexas, como o domínio de linguagens, a compreensão de fenômenos, o enfrentamento de situações-problema, a construção de argumentações e a elaboração de propostas. Essas competências devem ser desenvolvidas ao longo da escolaridade básica e são fundamentais para a formação dos futuros universitários, que deverão aplicá-las na geração de novos conhecimentos científicos, na proposição de soluções para problemas sociais e na promoção de inovações tecnológicas.

O ENEM utiliza princípios da Matemática Aplicada, da Estatística Avançada, da Informática e da Psicologia Cognitiva para fundamentar suas provas. Os itens das provas são elaborados com base em procedimentos pedagógicos rigorosos e analisados quantitativamente e qualitativamente, a fim de garantir sua qualidade. As análises quantitativas são baseadas na Teoria da Resposta ao Item (TRI). Os itens do ENEM são projetados para avaliar habilidades e competências, independentemente do domínio de conteúdo formal pelo estudante. As questões apresentam informações que permitem aos candidatos interpretar, inferir, deduzir e resolver problemas com base no conjunto de dados fornecidos. Esse modelo se afasta do foco exclusivo no conhecimento dos conteúdos escolares, como acontece nos vestibulares tradicionais. O objetivo é avaliar o que o aluno é capaz de fazer com as informações apresentadas, incentivando o desenvolvimento de competências mais amplas.

Em suma, o ENEM é um instrumento importante para avaliar as competências e habilidades dos estudantes ao final da educação básica, além de ser um instrumento

fundamental para o acesso ao ensino superior no Brasil. A partir de sua reformulação, o exame passou a ter um papel central nas oportunidades educacionais, contribuindo para a mobilidade acadêmica e para a reestruturação dos currículos do ensino médio.

RESULTADOS E DISCUSSÃO

Como elaborar os testes de múltipla escolha

O formato de múltipla escolha é amplamente utilizado em testes psicológicos e educacionais contemporâneos, mas sua construção exige um cuidado mais meticuloso do que outros formatos, devido à complexidade envolvida na formulação do enunciado e das alternativas de resposta. Para a elaboração desses itens, existem diretrizes que visam garantir uma construção sistemática e válida dos testes. Entre essas diretrizes, destaca-se o trabalho de Haladyna, Downing e Rodriguez (2002) que sintetizaram mais de 40 taxonomias acumuladas ao longo de 54 anos, consolidando 36 diretrizes.

Apesar das contribuições significativas dessas diretrizes, algumas dificuldades foram identificadas, como sobreposições e duplicações de conteúdo, linguagem imprecisa e um número excessivo de diretrizes. Para superar essas limitações, Moreno, Martínez e Muñiz (2004) desenvolveram um novo conjunto mais eficiente, composto por 12 diretrizes. Esse novo conjunto busca aumentar a validade dos instrumentos de avaliação, enfatizando a congruência com o propósito da avaliação. As diretrizes foram reorganizadas e reduzidas em número, eliminando redundâncias e focando em aspectos centrais.

Entre as 12 diretrizes propostas, duas estão relacionadas ao conteúdo a ser avaliado: (1) o conteúdo deve representar uma amostra do conteúdo especificado na tabela de especificação, evitando itens triviais; e (2) a representatividade deve guiar o nível de complexidade, especificidade, e a base de memorização ou raciocínio do item, bem como a melhor forma de expressá-lo. Outras três diretrizes tratam da expressão do conteúdo nos itens: (3) o ponto principal deve estar expresso na declaração, e cada opção deve concordar gramaticalmente com a declaração; (4) a sintaxe e a estrutura gramatical devem ser corretas, evitando ambiguidades, confusões e o uso excessivo de expressões negativas; e (5) a semântica deve corresponder ao conteúdo e ao público avaliado. As demais diretrizes são voltadas para a construção das opções de resposta: (6) deve haver apenas uma opção correta, acompanhada de alternativas plausíveis; (7) a opção correta deve estar distribuída em diferentes posições; (8) o número mínimo de

opções é três; (9) as opções devem ser apresentadas verticalmente; (10) o conjunto de opções para cada item deve parecer estruturado; (11) as opções devem ser autônomas, sem sobreposições ou referências a outras; e (12) nenhuma opção deve se destacar do restante em termos de conteúdo ou aparência Moreno, Martínez e Muñiz (2004).

A elaboração de questões eficazes requer atenção a princípios específicos que garantam a validade e a confiabilidade da avaliação (Haladyna, Downing e Rodriguez, 2002). Para que uma questão de múltipla escolha seja considerada eficaz, é necessário compreender os seus principais componentes: o enunciado, que introduz a situação ou problema a ser resolvido; as alternativas, que oferecem um conjunto de possíveis respostas, incluindo a resposta correta e os distratores (respostas incorretas, mas plausíveis); e a chave de resposta, que identifica a alternativa correta.

Ao desenvolver uma questão de múltipla escolha no contexto educacional, é fundamental que ela esteja alinhada aos objetivos de aprendizagem de uma disciplina ou de um ciclo de estudos. Além disso, tanto o enunciado quanto as alternativas devem ser claros e concisos, evitando ambiguidades ou o uso de jargões que possam confundir os estudantes (Trevisan e Amaral, 2016). Outro ponto importante é a utilização da Taxonomia de Bloom, que recomenda a formulação de questões que não apenas avaliem a memorização e a compreensão, mas que também abordem níveis mais elevados de cognição, como aplicação, análise, síntese e avaliação.

Para ilustrar, considere uma questão elaborada com o enunciado: "Considere o gráfico abaixo (não apresentado aqui), que mostra a função $f(t) = at^2 + bt + c$ representando a trajetória de um projétil lançado no plano cartesiano. O projétil atinge seu ponto mais alto no instante $t = 3$ e toca o solo novamente em $t = 6$. Sabendo que o gráfico corta o eixo y em $(0,2)$, determine o valor do coeficiente a". **Alternativas:** $A) -1/9;\ B) -2/9;\ C) -1/6;\ D) -2/6;\ E) -4/9$. **Resposta Correta:** $A) -1/9$. Esta questão requer que o estudante faça uma análise completa da equação de uma parábola e aplique conceitos de função quadrática, como vértice, raízes e a posição do gráfico no plano cartesiano. O aluno precisa interpretar os dados fornecidos no gráfico, identificar que o vértice da parábola ocorre em $t = 3$, que a raiz se dá em $t = 6$, e que o ponto de interseção com o eixo y é $(0,2)$.

O raciocínio começa com a identificação do formato da função quadrática: $f(x) = a(t-3)^2 + y_v$, onde y_v é a altura do vértice. Sabendo que a função corta o eixo y em $(0,2)$, o aluno precisa substituir $t = 0$ e $f(0) = 2$ para encontrar a relação

entre os coeficientes. Além disso, o estudante deve usar a condição de que a função zera em $t = 6$ para calcular o coeficiente a. Esse tipo de questão é um exemplo da etapa de análise e síntese da Taxonomia de Bloom, pois o aluno deve decompor o problema em suas partes componentes (análise) e reestruturar as informações para chegar a uma solução (síntese). Ele deve aplicar o conhecimento sobre funções quadráticas de maneira integrada e usar habilidades avançadas de cálculo e raciocínio lógico para resolver a questão.

A elaboração de enunciados eficazes é uma das etapas cruciais na criação de questões de múltipla escolha. Devem-se evitar negativas duplas, pois frases com negações complicadas tendem a confundir o leitor, o que pode comprometer a validade da questão. É preferível utilizar enunciados afirmativos, que são mais diretos e de fácil compreensão. Além disso, a contextualização das questões em situações relevantes para os estudantes pode melhorar o engajamento e a pertinência da avaliação, desde que o contexto seja relevante ao conteúdo avaliado (Haladyna, Downing e Rodriguez, 2002).

No desenvolvimento das alternativas, é recomendável utilizar entre quatro e cinco opções, embora a qualidade das alternativas seja mais importante que a quantidade. Os distratores, ou alternativas incorretas, devem ser plausíveis para que um aluno com entendimento superficial do conteúdo as considere como possíveis respostas corretas. Isso é crucial para que a questão consiga discriminar de forma eficaz os diferentes níveis de compreensão entre os estudantes. Também é importante evitar padrões previsíveis na posição da resposta correta, como sempre colocá-la em única opção. A alternância da posição da resposta correta ajuda a garantir que os participantes do teste não respondam baseados em suposições ou estratégias de adivinhação.

A análise e revisão das questões são etapas indispensáveis para garantir a validade e confiabilidade do teste. Cada questão deve ser revisada para assegurar que ela realmente mede o que se propõe a medir, sem influências de fatores irrelevantes. A revisão por pares é uma prática recomendada, pois permite que outros educadores identifiquem possíveis ambiguidades ou problemas de interpretação que o autor original possa não ter percebido (Haladyna, Downing e Rodriguez, 2002). Após a aplicação do teste, a análise de desempenho das questões deve ser realizada para identificar questões que apresentaram comportamento inesperado, como todos os estudantes acertando ou errando, o que pode indicar problemas na formulação da questão.

Considerações éticas também devem ser levadas em conta na formulação de questões de múltipla escolha. É essencial que as questões sejam formuladas de maneira

imparcial, sem privilegiar ou prejudicar nenhum grupo regional ou cultural. Isso inclui atenção à linguagem e ao contexto utilizado nas questões (Haladyna, Downing e Rodriguez, 2002). Além disso, ao utilizar cenários baseados em casos reais, deve-se ter cuidado para modificar detalhes que possam revelar informações confidenciais, preservando assim a privacidade e a ética na avaliação.

Elaborar questões de múltipla escolha que sejam válidas, confiáveis e justas é um desafio que requer prática e revisão contínua. Seguindo essas orientações, os educadores podem criar avaliações que realmente meçam as habilidades e conhecimentos dos alunos, contribuindo para um processo de ensino-aprendizagem mais eficaz.

TEORIAS PSICOMÉTRICAS

Para garantir a eficácia dos testes, é necessário compreender as teorias que fundamentam sua construção e análise. Embora a evolução dos testes tenha sido abordada historicamente, a importância das teorias que possibilitam sua construção muitas vezes não é totalmente reconhecida. As teorias estatísticas dos testes ajudam a estimar propriedades psicométricas essenciais, como a confiabilidade e validade dos testes, garantindo que as decisões baseadas nesses instrumentos sejam apropriadas e precisas. Sem essas teorias, seria impossível garantir a precisão e a relevância dos testes, o que poderia prejudicar significativamente os indivíduos avaliados.

A construção de testes é guiada por teorias psicométricas que orientam sua elaboração. A Teoria Clássica dos Testes (TCT) e a Teoria de Resposta ao Item (TRI) são duas abordagens principais nesse campo. A TCT, que se desenvolveu a partir dos trabalhos pioneiros de Charles Spearman no início do século XX, baseia-se na ideia de que a pontuação obtida em um teste é composta por uma pontuação verdadeira e um erro de medição (Muñiz, 2010). O modelo de Spearman assume que a pontuação verdadeira é a média das pontuações obtidas em uma infinidade de aplicações do teste e que o erro de medição é independente da pontuação verdadeira e entre diferentes aplicações do teste. Esses pressupostos são fundamentais para estimar a confiabilidade e validade dos testes, e formulam a base para a psicometria clássica.

Na Física, o conceito de escala espacial refere-se ao tamanho dos objetos com os quais interagimos, variando desde micrometros até centenas de metros (Andriola, 2009). Este conceito é fundamental para a compreensão da relação entre os mundos

microscópico e macroscópico. Avanços científicos recentes têm permitido uma melhor compreensão dessa relação. Por exemplo, o volume de um balão, que é uma propriedade macroscópica, depende das características microscópicas das partículas do gás que o infla. Entender a extensão das influências ao longo das escalas espaciais é crucial para a análise de problemas físicos e de outras naturezas.

De forma análoga, na Psicologia e na Educação, a Psicometria Moderna utiliza as duas abordagens para avaliação de testes. A Teoria Clássica dos Testes (TCT) concentra-se nas propriedades do teste como um todo, utilizando uma escala macroscópica. A Teoria da Resposta ao Item (TRI), por sua vez, foca no estudo dos itens individuais para realizar medições precisas, empregando uma abordagem microscópica. A principal diferença entre TRI e TCT reside na abordagem: a TRI analisa os itens individualmente, enquanto a TCT considera o teste como uma unidade. Ambas são válidas para avaliação de resultados de testes, cada uma com suas qualidades e especificidades.

Teoria da Resposta ao Item (TRI)

A TCT é amplamente utilizada e se consolidou ao longo dos anos com fórmulas e métodos estabelecidos para avaliar a precisão dos testes. No entanto, a abordagem clássica possui limitações. Ela não garante a invariância das medições em relação ao instrumento utilizado e depende das amostras para estimar propriedades psicométricas, o que pode comprometer a precisão dos resultados (Andriola, 2009).. Esses problemas levaram ao desenvolvimento da TRI, que surgiu para resolver essas questões e proporcionar uma medição mais robusta e uniforme.

A TRI, que começou a ser formalizada a partir dos anos 1960, introduz modelos mais complexos e detalhados, como o modelo logístico de Rasch. Essa teoria resolve algumas limitações da TCT ao garantir a invariância das estimativas das habilidades dos indivíduos e das propriedades dos itens (dificuldade e discriminação) independentemente do instrumento ou da amostra utilizada. A TRI proporciona uma abordagem mais refinada para medir variáveis psicológicas, permitindo uma estimativa mais precisa da confiabilidade e validade dos testes em diferentes contextos e para diferentes populações.

A TRI representa um avanço significativo na psicometria, oferecendo novos conceitos e ferramentas que complementam e expandem os métodos da TCT. No entanto, a TCT ainda é amplamente utilizada e continua a ser relevante na prática

psicológica. Ambas as abordagens coexistem e são aplicadas conforme a situação e as necessidades específicas da avaliação psicológica.

Os modelos teóricos não surgem de forma repentina ou linear. A TRI, apesar de resolver problemas da TCT, não contradiz suas conclusões fundamentais, mas adiciona suposições que permitem responder a questões não abordadas pela TCT. O desenvolvimento da TRI foi gradual, como revisado por Andriola (2009), começando há mais de 50 anos. L. L. Thurstone, em 1925, apresentou curvas associando idade e proporção de acertos, precursoras das modernas curvas características dos itens. M. W. Richardson, em 1936, formulou a função de informação do item. G. A. Ferguson, em 1942, desenvolveu conceitos equivalentes ao parâmetro de dificuldade da TRI. P. F. Lazarsfeld introduziu o conceito de traço latente em 1950.

F. M. Lord, em 1952, formulou os principais conceitos da TRI. Entre 1957 e 1958, A. Birnbaum propôs modelos logísticos com maior facilidade matemática. G. Rasch, em 1960, introduziu o modelo logístico de um parâmetro, destacando os princípios de convergência e separabilidade. Em 1968, F. M. Lord e M. R. Novick publicaram um trabalho que marcou o fim da fase teórica da TRI. A aceitação mundial da TRI só ocorreu duas décadas depois, com a publicação de Lord em 1980 sobre aplicações práticas da TRI.

Os termos "atributo latente" e "*construct*" são usados para se referir às características psicossociais que são de interesse para a medição. Um *construct* pode ser definido como aquilo que se pretende medir (Wu e Adams, 2007). A definição do *construct* está diretamente relacionada às questões de validade. A validade refere-se à medida das inferências, feitas a partir dos escores de um teste, que correspondem ao *construct* definido. Por exemplo, em testes de realização em larga escala, como os testes de matemática, o *construct* pode ser baseado no currículo de matemática. A teoria da resposta ao item (TRI) ajuda a vincular o *construct* aos tipos de inferências que podem ser feitas a partir dos escores do teste.

Na Teoria da Resposta ao Item (TRI), o *construct* tem um significado específico. Enquanto na Teoria Clássica dos Testes (TCT) as inferências são feitas principalmente com base no escore total obtido em um teste, na TRI o foco é medir o nível de um atributo latente em cada indivíduo, um atributo que não pode ser diretamente observado. Por exemplo, se o construct a ser medido é a proficiência em geometria, os itens do teste são desenhados para avaliar diferentes aspectos desse conhecimento específico.

A TRI parte do princípio de que o nível do atributo latente é o que determina as respostas do indivíduo aos itens, e não o contrário. Isso difere de abordagens em que o escore total depende diretamente dos itens específicos utilizados. Por exemplo, em um teste de matemática composto por questões de álgebra e geometria, a média dos escores dessas questões não define diretamente o nível de habilidade do aluno em matemática. Na TRI, é possível alterar os itens do teste, desde que continuem a medir o mesmo construct, e ainda assim manter uma avaliação consistente do nível do atributo latente, como a proficiência matemática.

O modelo TRI apresentado é unidimensional, o que significa que um conjunto de itens mede um único atributo latente. Se um teste contém vários atributos latentes, ele é multidimensional. Na prática, é difícil encontrar itens que testem exatamente o mesmo *construct*. Por exemplo, a solução de problemas matemáticos pode envolver habilidades de leitura e matemática, tornando o teste multidimensional.

Principais modelos logísticos da TRI

O modelo mais simples para a TRI é o logístico de um parâmetro, proposto por G. Rasch em 1960. Este modelo calcula a probabilidade de acerto de um item com base em sua dificuldade (parâmetro b). O modelo logístico de dois parâmetros, desenvolvido por A. Birnbaum entre 1957 e 1968, inclui a discriminação (parâmetro a), além da dificuldade (b). O modelo de três parâmetros, também de Birnbaum, adiciona um parâmetro para a chance de acerto ao acaso (c). Finalmente, o modelo de quatro parâmetros, proposto por Barton e Lord em 1981, inclui um parâmetro adicional (Yi) para situações em que indivíduos com alta competência não acertam o item devido a circunstâncias específicas (Wu e Adams, 2007).

O suposto da unidimensionalidade na TRI postula que a complexidade da resolução de um problema deve ser causada por uma única variável latente (θ). A unidimensionalidade sugere uma relação funcional entre θ e as respostas aos itens, representada pela curva característica do item. A independência local dos itens afirma que a resposta a um item não deve influenciar as respostas aos demais itens, o que é matematicamente expresso como o produto das probabilidades de acerto de cada item.

Fundamentos do Modelo de Rasch

O modelo de Rasch se fundamenta em dois pressupostos principais. Primeiro, assume que o atributo a ser medido pode ser representado em uma única dimensão, na

qual tanto as pessoas quanto os itens estão posicionados. Segundo Prieto (2003), a probabilidade de uma pessoa responder corretamente a um item é determinada pela diferença entre o nível de habilidade da pessoa e a dificuldade do item. O modelo usa uma função logística para modelar essa relação, onde a probabilidade de resposta correta é expressa pela fórmula:

$$ln\left(\frac{P_{is}}{1-P_{is}}\right) = \theta_s - \beta_i$$

Essa equação indica que o logaritmo da razão entre a probabilidade (P_{is}) do indivíduo *i* acertar o item *s* e a probabilidade de erro ($1 - P_{is}$) em um item é uma função da diferença entre o nível de habilidade da pessoa (θ_s) e a dificuldade do item (β_i). O modelo de Rasch prediz que a probabilidade de uma resposta correta será 50% quando a habilidade da pessoa é igual à dificuldade do item. Quando a habilidade é maior, a probabilidade de acerto aumenta, e quando é menor, a probabilidade diminui.

A expressão para a probabilidade de acerto no modelo de Rasch é uma função exponencial da forma sigmoide:

$$P_{is} = \frac{e^{(\theta_s - \beta_i)}}{1 + e^{(\theta_s - \beta_i)}}$$

A escala mais utilizada para interpretar os parâmetros é a escala *logit*, que é o logaritmo natural da razão de probabilidades, oferecendo uma interpretação intuitiva dos parâmetros dos itens e do estudante. Para aplicar o modelo de Rasch, é necessário estimar os parâmetros dos itens e das pessoas. Quando os parâmetros dos itens são conhecidos, é possível estimar os parâmetros individuais, usando métodos de estimação condicional. Caso contrário, é usada a estimativa conjunta, onde tanto os parâmetros dos itens quanto dos indivíduos são estimados simultaneamente. O método mais comum para essa estimativa é o de máxima verossimilhança, que busca os parâmetros que tornam as respostas observadas mais prováveis.

A precisão das estimativas é medida pelo erro típico de medida, que é inversamente proporcional à função de informação do teste. A função de informação é a soma das funções de informação dos itens, e a informação de um item varia ao longo do contínuo de habilidade, sendo máxima quando a dificuldade do item corresponde ao nível de habilidade do indivíduo.

O modelo de Rasch apresenta algumas vantagens em relação a outros modelos aplicados na Teoria de Resposta ao Item (TRI). Destacam-se as seguintes características:

Medida conjunta: Os parâmetros dos indivíduos e dos itens são expressos nas mesmas unidades e localizados no mesmo contínuo, permitindo uma análise mais realista e detalhada. A interpretação das pontuações não depende das normas de grupo, mas sim da probabilidade de uma pessoa resolver corretamente determinados itens.

Objetividade específica: As comparações entre indivíduos e itens são independentes o que garante que a diferença entre duas pessoas ou entre dois itens não dependa das características da amostra.

Propriedades de intervalo: A escala *logit* possui propriedades de intervalo, o que permite utilizar análises paramétricas rigorosas, garantindo a invariança das pontuações diferenciais ao longo do contínuo.

Especificidade do erro típico de medida: O modelo de Rasch permite quantificar a informação fornecida por um teste em diferentes pontos do contínuo e ajustar a dificuldade dos itens para aumentar a precisão das medições.

Para garantir que as vantagens do modelo de Rasch sejam efetivas, é essencial que os dados se ajustem ao modelo. A presença de respostas que se afastam do esperado pode indicar que os parâmetros dos sujeitos e dos itens não têm significado teórico. A análise de ajuste pode ser realizada usando análise estatísticas baseadas em resíduos, que ajudam a identificar itens e indivíduos que não se ajustam ao modelo.

CONSIDERAÇÕES FINAIS

Considerando a revisão da literatura sobre testes padronizados e suas aplicações, especialmente no contexto educacional, é possível concluir que estes instrumentos desempenham um papel crucial na avaliação do desempenho acadêmico e na tomada de decisões educacionais. No entanto, sua utilização não está isenta de críticas e desafios. A objetividade, a padronização e a capacidade de comparar resultados entre diferentes grupos são as principais vantagens dos testes padronizados. Eles fornecem dados quantitativos que podem ser utilizados para identificar áreas de domínio e deficiência dos estudantes, informando a prática pedagógica e a elaboração de políticas educacionais.

Por outro lado, os testes padronizados também apresentam limitações. A ênfase em habilidades específicas e a natureza objetiva das questões podem restringir a avaliação de habilidades mais complexas, como a criatividade e o pensamento crítico. Além disso, a validade e a confiabilidade dos testes podem ser questionadas, especialmente quando são utilizados para tomar decisões de alta relevância, como a seleção para programas de ensino superior.

A equidade na aplicação dos testes padronizados é outro ponto crucial a ser considerado. Fatores como o contexto socioeconômico, cultural e linguístico dos estudantes podem influenciar o desempenho nos testes, levando a resultados injustos. É fundamental que os testes sejam adaptados culturalmente e que sejam consideradas as diferentes realidades dos estudantes ao interpretar os resultados.

A evolução histórica dos testes padronizados demonstra a busca constante por instrumentos de avaliação mais precisos e justos. No entanto, ainda há muito a ser feito para garantir que esses testes sejam utilizados de forma adequada e ética. É necessário investir em pesquisas que investiguem os impactos dos testes padronizados no longo prazo, bem como em práticas de avaliação que complementem os testes tradicionais, como a avaliação formativa e a avaliação por portfólio.

A Teoria da resposta ao item (TRI) trouxe mudanças significativas para a avaliação psicológica e educacional, focando a análise nos itens individuais em vez do teste como um todo. Modelos unidimensionais de TRI são amplamente utilizados, mas a TRI multidimensional, que lida com dados politômicos, está se tornando cada vez mais relevante. No Brasil, a TRI é utilizada por instituições como o Inep e algumas secretarias estaduais de educação. No entanto, há necessidade de mais formação e pesquisa na área. A evolução futura pode levar ao uso mais significativo dos modelos multidimensionais de TRI, refletindo a tendência de que o uso massivo de modelos unidimensionais está em declínio.

Com base nesta revisão da literatura, conclui-se que os testes padronizados são ferramentas valiosas, mas que devem ser utilizados com cuidado e complementados por outras formas de avaliação. A busca por uma avaliação mais justa e completa do desempenho dos estudantes continua sendo um desafio para a comunidade educacional. Em suma, os instrumentos avaliativos padronizados são ferramentas importantes para a avaliação educacional, mas devem ser utilizados com cautela e em conjunto com outras formas de avaliação. É fundamental que os educadores e pesquisadores e estejam

atentos às suas limitações e trabalhem para garantir que os testes sejam utilizados de forma a promover a equidade e a justiça na educação.

As seguintes sugestões podem ser desenvolvidas em futuras pesquisas: investigar o impacto da utilização de diferentes tipos de testes padronizados no desempenho dos estudantes; Analisar a relação entre o desempenho em testes padronizados e outros indicadores de sucesso acadêmico, como a frequência escolar e a participação em atividades extracurriculares; Desenvolver instrumentos de avaliação que combinem elementos dos testes padronizados com outras formas de avaliação, como a avaliação formativa e a avaliação por portfólio; Investigar o impacto da cultura e do contexto socioeconômico no desempenho em testes padronizados.

REFERENCIAS

ANDRIOLA, W. B. Doze motivos favoráveis à adoção do Exame Nacional do Ensino Médio (ENEM) pelas instituições federais de ensino superior (IFES). Ensaio: avaliação e políticas públicas em educação, v. 19, n. 70, p. 107-125, 2011.

ANDRIOLA, W.B. Psicometria moderna: Características e tendências. Estudos em Avaliação Educacional, 20, n. 43, 319-340, 2009.

FORZANI, E. et al. Advances and missed opportunities in the development of the 2026 NAEP Reading Framework. Literacy Research: Theory, Method, and Practice, v. 71, n. 1, p. 153-189, 2022.

GREGORY, R. J. Psychological testing: History, principles and applications (6nd ed.). New Jersey: Prentice-Hall, Inc, 2010.

HALADYNA, Thomas M.; DOWNING, Steven M.; RODRIGUEZ, Michael C. A review of multiple-choice item-writing guidelines for classroom assessment. Applied measurement in education, v. 15, n. 3, p. 309-333, 2002.

HALADYNA, T. M.; RODRIGUEZ, M. C. A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. Educational Measurement: Issues and Practice, v. 32, n. 1, p. 14-24, 2013.

MONTEIRO, A. F. Teste de escolha múltipla: construção, aplicação e avaliação. Revista Brasileira de Educação, v. 15, n. 44, p. 324-341, 2010.

MORENO, Rafael; MARTÍNEZ, Rafael J.; MUÑIZ, José. Directrices para la construcción de ítems de elección múltiple. Psicothema, v. 16, n. 3, p. 490-497, 2004.

MUÑIZ, J. Las teorias de los test: teoría clásica y teoría de respuesta a los ítems. Papeles del Psicólogo, v. 31, n. 1, 57-66, 2010.

NAEP - National Assessement of Educational Progress. Mathematicas framework for the 2026. Washington, DC: National Assessment Governing Board (U.S. Department of Education), 2021.

TREVISAN, André Luis; AMARAL, Roseli Gall do. A Taxionomia revisada de Bloom aplicada à avaliação: um estudo de provas escritas de Matemática. Ciência & Educação (Bauru), v. 22, n. 2, p. 451-464, 2016.

PRIETO, G.; DELGADO, A. Análisis de un test mediante el modelo de Rasch. Psicothema, v. 15, n. 1, 94-100, 2003.

URBINA, S. Essentials of psychological testing. New Jersey: John Wiley & Sons, Inc, 2004.

WU, M., & ADAMS, R. Applying the Rasch model to psycho-social measurement: A practical approach. Melbourn: Educational Measurement Solutions, 2007.