UNIVERSIDADE DE BRASÍLIA – UNB

FACULDADE DE EDUCAÇÃO – FE

PROGRAMA DE PÓS-GRADUAÇÃO EM EDUCAÇÃO MODALIDADE PROFISSIONAL – PPGE-MP

O ERRO COMO FERRAMENTA PARA COMPREENDER O DESEMPENHO DE ESTUDANTES EM AVALIAÇÕES EM LARGA ESCALA

DISSERTAÇÃO DE MESTRADO

LAÍS SILVEIRA ANTONIETTO

BRASÍLIA/DF

SETEMBRO/2022

LAÍS SILVEIRA ANTONIETTO

O ERRO COMO FERRAMENTA PARA COMPREENDER O DESEMPENHO DE ESTUDANTES EM AVALIAÇÕES EM LARGA ESCALA

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Educação – Modalidade Profissional da Universidade de Brasília, vinculado à linha de pesquisa Políticas Públicas e Gestão da Educação Profissional e Tecnológica, como requisito para a obtenção do título de Mestre em Educação.

Orientadora: Dra. Claudia Maffini Griboski.

BRASÍLIA/DF

SETEMBRO/2022

LAÍS SILVEIRA ANTONIETTO

O ERRO COMO FERRAMENTA PARA COMPREENDER O DESEMPENHO DE ESTUDANTES EM AVALIAÇÕES EM LARGA ESCALA

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Educação – Modalidade Profissional da Universidade de Brasília, vinculado à linha de pesquisa Políticas Públicas e Gestão da Educação Profissional e Tecnológica, como requisito para a obtenção do título de Mestre em Educação.

Orientadora: Dra. Claudia Maffini Griboski.

COMISSÃO EXAMINADORA

_____

Professora Dra. Claudia Maffini Griboski
Orientadora – Universidade de Brasília (UnB)

_____

Professora Dra. Girlene Ribeiro de Jesus
Examinadora – Universidade de Brasília (UnB)

_____

Professor Dr. João Luiz Horta Neto
Examinador Externo – Inep

_____

Professor Dr. Francisco Thiago Silva
Examinador suplente – Universidade de Brasília (UnB)

AGRADECIMENTOS

Quero, primeiramente, agradecer ao meu parceiro de todos os dias, Maurício, que aguentou todos os estresses e que esteve ao meu lado, me pentelhando sempre que possível para que eu não desistisse. Esse trabalho também é seu.

Agradeço também aos meus pais, que sempre foram para mim exemplo e que sempre me instigaram a querer mais e alcançar mais.

Ao meu irmão Lucas, pelo apoio acadêmico e pela inspiração que sempre foi e é.

Ao meu sogro, sogra e cunhada, por todos os dias e noites que cuidaram do meu filho para que eu conseguisse concluir este trabalho.

À Carol e ao André, que, mesmo de longe, sempre arranjaram um jeito de se fazerem presentes.

À amiga Mari, que fez esse período da minha vida mais leve e feliz.

Aos meus ex-colegas de trabalho que, à época da admissão no mestrado, fizeram toda a diferença para que hoje eu estivesse aqui. Agradeço especialmente à minha ex-chefe, Elianice, que contribuiu enormemente para que conseguisse desenvolver esse trabalho.

Agradeço à professora Terezinha, minha mentora profissional e acadêmica, que deu início a tudo que aqui está escrito.

À Gabriela, pelo apoio nas análises que permitiram a conclusão deste trabalho.

À Prefeitura de São Paulo, especialmente ao Núcleo Técnico de Avaliação, pela autorização de acesso aos dados da Prova São Paulo 2018.

Ao Cebraspe, pelo apoio institucional e pela disponibilização dos dados estatísticos a esta pesquisa.

À minha orientadora, pela paciência e por sempre me acolher quando eu precisava.

À Capes/CNPq, pela disponibilização das bases de pesquisa.

Ao PPGE-MP, pelo suporte e acolhimento durante o período da pandemia.

Ao colega de mestrado Fábio, pela ajuda oferecida e pelo encorajamento.

À colega Danuzia, pela luz dada nas análises aos 45 do segundo tempo.

Eu dedico este trabalho aos meus filhos.

RESUMO

Esta dissertação tem por objetivo analisar os erros cometidos por estudantes em um teste cognitivo de avaliação em larga escala e, a partir dessa análise, gerar evidências pedagógicas para compreender o desempenho deles nessas avaliações. Neste trabalho, parte-se do pressuposto de que o desenvolvimento de uma habilidade é um processo contínuo, cujos estágios são evidenciados pelos erros que os estudantes cometem. Aplicando esse conceito ao universo das avaliações em larga escala e considerando que as opções erradas de um item refletem esses estágios, entende-se que estudantes em estágios diferentes de desenvolvimento de uma habilidade cometem erros também diferentes. Sendo assim, este estudo utilizou uma metodologia de análise estatística para analisar as probabilidades de marcação das opções erradas e associá-las ao nível de proficiência desses estudantes, criando uma escala interpretada que leva em consideração os erros dos estudantes ao invés dos acertos. A questão central que norteou essa investigação foi: é possível utilizar as marcações dos estudantes nas opções erradas para gerar evidências sobre o seu desempenho em avaliações em larga escala? Para responder a esse questionamento, a presente dissertação justifica o estudo com base em uma abordagem quali-quantitativa, sendo organizada na forma de artigo, com introdução, dois estudos analíticos, um produto técnico, que traz uma escala interpretada com o objetivo de oferecer à comunidade escolar e à sociedade uma devolutiva a respeito do desempenho dos estudantes em avaliações em larga escala e as considerações finais. Os resultados permitem propor a referida escala, delimitando para cada um dos grupos de estudantes estabelecidos um perfil dos erros geralmente cometidos. Cabe destacar que essa pesquisa tem caráter exploratório, tendo em vista seu ineditismo no cenário das análises que se fazem dos dados extraídos em avaliações em larga escala. Admite-se, portanto, suas limitações no que se refere aos modelos estatísticos empregados, cabendo análises mais robustas para consolidar os procedimentos adotados. Por outro lado, destaca-se como grande contribuição deste trabalho uma análise que complementa as escalas de proficiência tradicionais e se soma aos resultados das avaliações em larga escala, oferecendo uma nova perspectiva sobre o desempenho dos estudantes.

Palavras-chave: avaliação em larga escala; escala interpretada de proficiência; análise do erro na aprendizagem

ABSTRACT

This dissertation aims to analyze the errors made by students in a large-scale cognitive assessment test and, from this analysis, generate pedagogical evidence to understand their performance in these assessments. In this work, it is assumed that the development of a skill is a continuous process, which stages are evidenced by the mistakes students make. Applying this concept to the universe of large-scale assessments and considering that the wrong choices of an item reflect these stages, it is understood that students at different stages of development of a skill also make different mistakes. Therefore, this study used a statistical analysis methodology to analyze the probabilities of marking the wrong options and associate them with the proficiency level of these students, creating an interpreted scale that considers the students' mistakes instead of correct markings. The central question that guided this investigation was: is it possible to use students' marks in the wrong options to generate evidence about their performance in large-scale assessments? To answer this question, the present dissertation justifies the study based on a quali-quantitative approach, being organized in the form of an article, with an introduction, two analytical studies, a technical product, which brings an interpreted scale with the objective of offering the school community and society feedback on student performance in large-scale assessments and final considerations. The results allow us to propose this scale, delimiting for each of the groups of students a profile of the errors generally made. It is worth noting that this research has an exploratory character, given its novelty in the scenario of analyzes performed on data extracted in large-scale evaluations. Therefore, its limitations regarding the statistical models used are admitted, requiring more robust analyzes to consolidate the adopted procedures. On the other hand, an analysis that complements traditional proficiency scales and adds to the results of large-scale assessments stands out as a major contribution of this work, offering a new perspective on student performance.

Keywords: large-scale assessment; interpreted scale of proficiency; error analysis in learning

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

Aneb	- Avaliação Nacional da Educação Básica
Anresc	- Avaliação Nacional do Rendimento Escolar
BNCC	- Base Nacional Comum Curricular
CCI	- Curva Característica do Item
CF	- Constituição Federal
GERES	- Estudo Longitudinal sobre a Qualidade e Equidade no Ensino Fundamental Brasileiro
Inep	- Instituto Nacional de Pesquisas Educacionais Anísio Teixeira
MEC	- Ministério da Educação
MRN	- Modelo de Respostas Nominais
MRG	- Modelo de Respostas Graduais
OCDE	- Organização para a Cooperação e Desenvolvimento Econômico
PPGE-MP	- Programa de Pós-Graduação em Educação - Modalidade Profissional
PUC Rio	- Pontifícia Universidade Católica do Rio de Janeiro
Saeb	- Sistema de Avaliação da Educação Básica
SME-SP	- Secretaria Municipal de Ensino de São Paulo
TCT	- Teoria Clássica dos Testes
TRI	- Teoria de Resposta ao Item
UFBA	- Universidade Federal da Bahia
UFJF	- Universidade Federal de Juiz de Fora
UFMG	- Universidade Federal de Minas Gerais
Unicamp	- Universidade de Campinas

SUMÁRIO

INTRODUÇÃO

Saber onde se pretende chegar e como se quer chegar pressupõe o conhecimento do lugar onde se está e de como se está preparado para chegar. Na educação, a lógica não é diferente. Se queremos uma educação de qualidade e acessível a todos, segundo os princípios garantidos pela Constituição Federal (CF/1988), se faz necessário investigar a realidade das escolas brasileiras, em seus diversos aspectos e contextos, para que se possa trilhar os caminhos que levarão ao alcance desse objetivo.

Nessa investigação se inserem os estudos em avaliação educacional, que permitem averiguar de que maneira as avaliações estão sendo conduzidas, quais são os seus impactos e como podem contribuir para um ensino de qualidade. Klein e Fontanive (1995, p. 29) afirmam que "a avaliação educacional é um sistema de informações que tem por objetivos fornecer diagnóstico e subsídios para a implementação ou manutenção de políticas educacionais". Sendo assim, consiste em uma ferramenta para indicar em que situação se encontra a educação, o que é primordial para se estabelecer novas práticas ou reforçar as que estejam dando certo. Sousa e Oliveira (2010, p. 801) vão na mesma direção quando postulam que,

> [...] ao realizar determinado processo avaliativo, espera-se, explicitamente, verificar quão distante se está da situação desejável e, a partir daí, definir elementos para modificar a situação em direção ao padrão desejado. Além disso, a avaliação pode, também, possibilitar a emergência de propostas de redirecionamento ou transformação da situação avaliada.

Partindo do pressuposto de que a avaliação pode "produzir um ensino de melhor qualidade" (SOUSA; OLIVEIRA, 2010 *apud* SOUSA, 2001, p. 90), a partir da década de 1990, observa-se a disseminação de avaliações por todo o país, nos diversos níveis e modalidades de ensino, que passam a ocupar espaço de destaque nas políticas públicas de educação. Conforme Sousa e Oliveira (2010, p. 794) colocam, a avaliação passou a ser

> [...] recomendada e promovida por agências internacionais, pelo Ministério da Educação e por Secretarias de Educação de numerosos estados brasileiros, como elemento privilegiado para a realização das expectativas de promoção da melhoria da qualidade do ensino básico e superior.

Esse movimento é seguido de um aumento na quantidade de estudos (BAUER, 2012) que têm como objetivo entender os pressupostos que fundamentam essas avaliações, os mecanismos implícitos na sua execução e as consequências que surgem delas para a comunidade escolar e a sociedade como um todo.

Destacam-se entre essas pesquisas, aquelas que se dedicam a investigar o uso dos resultados de avaliações educacionais e o seu impacto nas políticas públicas, na prática pedagógica, na gestão de escolas e na aprendizagem dos estudantes. Parte desses estudos relatam que, apesar dos esforços no sentido de dar transparência aos resultados das avaliações, muitos professores e gestores revelam desconhecê-los, não os utilizar ou até mesmo terem dificuldade para se apropriarem de conceitos básicos ou da metodologia empregada em avaliações (LONGO, 2019; KISTEMANN JR; GOUVEA, 2019; PERRY, 2009). Por outro lado, há estudos que evidenciam que, embora se constate um aprimoramento nas técnicas de avaliação e uma maior regularidade na sua execução, não se observa um incremento na indução de políticas ou nas tomadas de decisão a partir dos seus resultados (SOUSA; OLIVEIRA, 2010).

Como reflexo dessas investigações a discussão em torno da temática alimenta duas vertentes: a dos que alegam que a avaliação em larga escala traz impactos negativos ao estabelecer o ranqueamento de estudantes e escolas, desconsiderar aspectos característicos da região onde estão inseridos, incentivar o treinamento para testes, promover o reducionismo curricular, entre outros[1]; e a daqueles que defendem a avaliação como importante mecanismo de monitoramento da educação pública brasileira, que é evidenciada por inúmeros estudos. Um deles é o de Sousa e Ferreira (2019, p. 16), em que se argumenta que

[1] Em estudo publicado em 2019, Souza coleta dados de pesquisas que usam como fonte de informação discursos de educadores sobre avaliação em larga escala. Como resultado, ele identificou que, apesar de representarem o sentimento dos educadores, muitas das críticas desses profissionais sobre a temática são infundadas e advêm do senso comum que se faz crescer diante da pouca informação que possuem sobre avaliação (SOUZA, 2019).

> [...] desconsiderar a relevância da avaliação de larga escala, sobretudo no Brasil, seria leviandade e injustiça, principalmente porque ela representa não só aqui, mas em muitos países, uma ação eficaz de reestruturação da Escola e do Sistema de Educação, capaz de definir critérios essenciais pelos quais se deve compreender a qualidade do trabalho educacional. Os indicadores de qualidade produzidos pelas análises dessas avaliações permitem compreender o desempenho do aluno, associado às contingências sociais, à estrutura e às condições da escola que definem o bom desempenho. Revelam também como a formação do professor está relacionada com o rendimento do aluno e como o nível socioeconômico da clientela escolar é decisivo no desempenho acadêmico individual.

É nesse cenário conflituoso de avaliações cada vez mais presentes no cotidiano escolar, porém, sem a adequada apreensão dos seus princípios e das suas possibilidades de uso, que se insere esta pesquisa. Considerando a relevância que as avaliações alcançaram no cenário nacional, redefinindo e reorientando políticas públicas de grande impacto no sistema educacional brasileiro, as pesquisas sobre essa temática se tornaram fundamentais para melhor entender seus procedimentos e resultados. Como afirmam Jesus, Rêgo e Souza (2018, p. 860),

> [...] testes educacionais são aplicados anualmente para milhões de estudantes da educação básica e milhares da educação superior no Brasil. Os resultados obtidos pelos estudantes são utilizados para ingresso no ensino superior, responsabilização dos sistemas de ensino, das unidades escolares e das instituições de ensino superior. Além disso, os escores obtidos nos testes servem para o cálculo de indicadores e para a indução de políticas públicas educacionais.

Por isso, cabe à pesquisa científica a tarefa de buscar entender os mecanismos pertinentes a essas avaliações, de modo a garantir que sejam de fato utilizadas em prol da educação e do trabalho escolar, e não como alavanca política ou instrumento de exclusão social. O presente estudo visa cumprir com essa tarefa, considerando esse lugar de destaque que a avaliação da educação ganhou nas discussões acadêmicas, nos discursos políticos e na mídia e o impacto que têm sobre as políticas públicas nacionais em educação.

Nesse propósito, esta pesquisa pretende oferecer à comunidade acadêmica e à sociedade em geral um estudo sistematizado sobre avaliação educacional, a fim de esclarecer seus fundamentos e propósitos, visando auxiliar na compreensão do desempenho dos estudantes a partir de formas alternativas de interpretação dos seus resultados. Assim, espera-se superar o **problema** que deu origem ao presente estudo, a saber: **(i) os resultados das avaliações em larga escala**

atualmente pouco auxiliam na compreensão do desempenho dos estudantes; e **(ii) muitas vezes, esses resultados não são suficientes para gerar evidências passíveis de serem utilizadas para intervenções pedagógicas na escola.**

O trabalho aqui descrito é fruto dessa inquietação, também vivenciada na instituição em que trabalho, responsável, entre outras funções, pela realização de avaliações educacionais em larga escala no Brasil. Apesar dos esforços empregados no sentido de tornar didática e compreensível a imensa gama de dados que esse tipo de avaliação é capaz de gerar, na prática, ainda é considerável o quantitativo de relatos que evidenciam a dificuldade dos professores de diversas regiões do país em compreendê-los e utilizá-los em sala de aula.

Por outro lado, como alguém que trabalha com avaliações educacionais e é acostumada a lidar com os seus resultados, é igualmente inquietante constatar que existe ainda um dado que tem passado despercebido pelas análises e relatórios de qualquer instituição no país que se dedique a tal: o erro. Nas técnicas mais modernas de elaboração de itens de avaliação educacional, que, majoritariamente, utilizam itens de múltipla-escolha, as opções erradas pretendem representar os possíveis erros que estudantes cometeriam no processo de desenvolvimento da habilidade que está sendo medida naquele item (DOWNING; HALADYNA, 2006; ZIMMARO, 2016).

Pensando nessa lógica, emerge a **pergunta central que orienta esta pesquisa:** *é possível utilizar as marcações dos estudantes nessas opções erradas para gerar evidências sobre o seu desempenho em avaliações em larga escala?* A essa indagação se sucedem outras duas: *qual método permitiria a compilação e o tratamento estatístico dessas informações? Esses resultados poderiam ser apresentados ao público no formato de uma escala interpretada?*

Tendo essa investigação como norte, o primeiro passo da presente dissertação foi buscar entender como atualmente são geradas evidências sobre o desempenho dos estudantes em avaliações educacionais no Brasil. Para tanto, foi selecionada uma avaliação de caráter nacional, referência para diversas outras iniciativas de avaliação no país, o Sistema de Avaliação da Educação Básica (Saeb).

No início da década de 1990, o Estado Brasileiro definiu e implementou o Saeb como uma iniciativa de avaliação educacional em larga escala direcionada aos estudantes da educação básica (ensinos fundamental e médio). Em consonância com a Lei de Diretrizes e Bases da Educação Nacional (Lei n. 9.394/1996), o objetivo dessa avaliação é garantir o monitoramento do

rendimento escolar na educação básica ofertada nas redes públicas de ensino do país, "objetivando a definição de prioridades e a melhoria da qualidade do ensino" (BRASIL, 1996). Ademais, o Saeb busca atender à meta 7 do Plano Nacional de Educação (Lei n. 13.005/2014), fomentando "a qualidade da educação básica em todas as etapas e modalidades, com melhoria [...] da aprendizagem de modo a atingir as seguintes médias nacionais para o Ideb" (BRASIL, 2014). Mais especificamente, o Saeb foi institucionalizado com a finalidade de

> [...] (i) desenvolver e aprofundar a capacidade avaliativa das unidades gestoras do sistema educacional (MEC, secretarias estaduais e órgãos municipais); regionalizar a operacionalização do processo avaliativo, criando nexos e estímulos para o desenvolvimento de infraestrutura de pesquisa e avaliação educacional; propor uma estratégia de articulação dos resultados das pesquisas e avaliações já realizadas ou em vias de implementação (BRASIL/MEC/INEP, s.d., p. 3 *apud* BONAMINO; FRANCO, 1999, p. 111).

Conforme afirmam Bonamino e Franco (1999), dessa visão descentralizada e participativa do Saeb, se evoluiu, com o passar dos ciclos, para uma centralização da iniciativa no âmbito do Ministério da Educação (MEC). Inicialmente, a avaliação se dava com uma amostra de estudantes das 1ª, 3ª, 5ª e 7ª séries das redes públicas de ensino, cujas provas focavam as áreas de língua portuguesa, matemática e ciências. A partir de 1995, houve alterações no seu desenho metodológico e no público-alvo que permanecem até hoje, tais como:

> [...] i) inclusão da rede particular de ensino na amostra; ii) adoção da Teoria de Resposta ao Item (TRI), que permite estimar as habilidades dos alunos independentemente do conjunto específico de itens respondidos; iii) opção de trabalhar com as séries conclusivas de cada ciclo escolar (4ª e 8ª série do ensino fundamental e inclusão da 3ª série do ensino médio); iv) priorização das áreas de conhecimento de língua portuguesa (foco em leitura) e matemática (foco em resolução de problemas); v) participação das 27 unidades federais; vi) adoção de questionários para os alunos sobre características socioculturais e hábitos de estudo. A partir da introdução dessas inovações, o Saeb tornou comparáveis os desempenhos dos alunos entre anos e séries (BONAMINO; SOUSA, 2012, p. 376-377).

Com a entrada do milênio, outra grande inovação observada na trajetória do Saeb foi a ampliação do seu escopo, em 2005, com a publicação da portaria n. 931, que deu origem à Avaliação Nacional da Educação Básica (Aneb) e à Avaliação Nacional do Rendimento Escolar (Anresc), conhecida popularmente como Prova Brasil. A primeira manteve o caráter amostral,

incluindo escolas da rede pública e privada, mas a segunda passou a abranger todos os estudantes concluintes das séries finais do ensino fundamental I e II (4ª série/5º ano e 8ª série/9º ano) e do 3º ano do ensino médio da rede pública de ensino. Um dos objetivos dessa alteração era oferecer a divulgação de resultados por unidade escolar, de modo que elas pudessem se apropriar melhor da situação em que se encontravam e investigar estratégias específicas para a melhoria da qualidade do ensino que ofertavam. Esse novo formato se consolidou e constitui ainda hoje a base geral de atuação do Saeb, que, em 2019, teve como objetivos:

> [...] (i) avaliar a qualidade, a equidade e a eficiência da educação praticada no país em seus diversos níveis governamentais; (ii) produzir indicadores educacionais para o Brasil, suas regiões e unidades da federação e, quando possível, para os municípios e as instituições escolares, tendo em vista a manutenção da comparabilidade dos dados, permitindo, assim, o incremento das séries históricas; (iii) subsidiar a elaboração, o monitoramento e o aprimoramento de políticas públicas baseadas em evidências, com vistas ao desenvolvimento social e econômico do Brasil; e (vi) desenvolver competência técnica e científica na área de avaliação educacional, ativando o intercâmbio entre instituições educacionais de ensino e pesquisa (INEP, 2019, p. 9).

Entrando especificamente na questão das evidências de desempenho dos estudantes que o Saeb oferece como resultado de sua aplicação, a presente pesquisa optou por investigar um dos documentos que dele resultam e que é divulgado à sociedade como o conjunto de habilidades que os estudantes demonstraram ter desenvolvido ao responderem os testes. Esse documento recebe o nome de escala de proficiência e consiste em descrições feitas por especialistas de cada área avaliada a partir das respostas corretas dadas pelos estudantes aos itens aplicados, construindo um panorama dos diversos níveis de desempenho em certo componente curricular. Essa escala é viabilizada pela adoção da Teoria de Resposta ao Item (TRI)[2].

Como um dos principais instrumentos de devolutiva do Saeb – e de avaliações em larga escala em geral –, os níveis de proficiência têm papel fundamental na compreensão dos resultados dos estudantes e no estabelecimento de políticas que visem à qualidade do ensino. Tendo em vista tamanha importância, não é de se estranhar que pesquisas em educação façam de seu foco os pressupostos teóricos que embasam a construção desses documentos, a sua acepção pelos professores e o seu papel na definição de objetivos educacionais.

[2] A TRI abrange uma série de modelos que visam explicar, probabilisticamente, o desempenho de indivíduos perante um determinado teste (ARAUJO; ANDRADE; BORTOLOTTI, 2009).

Visando se encaixar nesse grupo e entendendo que, conforme afirma Klein (2006, p. 152), "a interpretação da escala e as informações sobre os erros dos alunos deveriam ser utilizadas para fornecer subsídios para programas de formação e capacitação de professores", no presente estudo foi feita uma busca em bases de dados de teses e dissertações e em plataformas de publicação de artigos acadêmicos, por pesquisas que investigassem escalas de proficiência, especialmente a do Saeb, e avaliações em larga escala, suas concepções e implicações nas últimas duas décadas. Utilizando os termos indutores avaliação educacional, avaliação em larga escala, Saeb e escala de proficiência, foi possível fazer um levantamento inicial, filtrado com base na relevância para a pesquisa composta neste trabalho.

Com esse levantamento, se espera tecer "uma rede de trabalhos e pesquisas ligados por categorias e sínteses do conhecimento que ganham significado quando são inventariados, ordenados, classificados e relacionados com o objeto que se esteja pesquisando" (SILVA; BORGES, 2018, p. 1694), desenvolvendo o estado do conhecimento da temática abordada na presente pesquisa. O Quadro 1 a seguir elenca os estudos selecionados, que serão em seguida discutidos conforme ordenação abaixo.

Quadro 1 - Trabalhos selecionados para a composição do estado do conhecimento do presente estudo (continua)

Trabalhos sobre avaliação educacional em larga escala e seus impactos		
João Luiz Horta Neto	Avaliação externa: a utilização dos resultados do Saeb 2003 na gestão do sistema público de ensino fundamental no Distrito Federal.	Dissertação (2006)
Sandra Zákia Sousa e Romualdo Portela de Oliveira	Sistemas estaduais de avaliação: uso dos resultados, implicações e tendências.	Artigo (2010)
Alicia Bonamino e Sandra Zákia Sousa	Três gerações de avaliação da educação básica no Brasil: interfaces com o currículo da/na escola.	Artigo (2012)
Adriana Bauer	Estudos sobre sistemas de avaliação educacional no Brasil: um retrato em preto e branco	Artigo (2012)
João Luiz Horta Neto	As avaliações externas e seus efeitos sobre as políticas educacionais: uma análise comparada entre a União e os estados de Minas Gerais e São Paulo.	Tese (2013)
Fátima Soares da Silva e Telma Ferraz Leal	Escala de proficiência da prova brasil: o que informa aos professores?	Artigo (2018)
Rodrigo Marques, Ronildo Stieg e Wagner dos Santos	Exames estandardizados: análise dos modelos e das teorias na produção acadêmica.	Artigo (2020)

Quadro 1 - Trabalhos selecionados para a composição do estado doo conhecimento do presente estudo (conclusão)

Trabalhos sobre metodologia de construção de escalas		
Ruben Klein	Utilização da Teoria de Resposta ao Item no Sistema Nacional de Avaliação da Educação Básica (Saeb).	Artigo (2003)
Lina Kátia Mesquita de Oliveira, Creso Franco e Tufi Machado Soares	Projeto GERES/2005: novos indicadores para construção e interpretação da escala de proficiência.	Artigo (2007)
Nilma Santos Fontanive, Ligia Gomes Elliot e Ruben Klein	Os desafios da apresentação dos resultados da avaliação de sistemas escolares a diferentes públicos.	Artigo (2007)
Lina Kátia Mesquita de Oliveira	Três investigações sobre escalas de proficiência e suas interpretações.	Tese (2008)
Flávia Almeida Perry	Escalas de proficiência: diferentes abordagens de interpretação na avaliação educacional em larga escala.	Dissertação (2009)
Trabalhos sobre o erro em avaliações em larga escala		
Maria José Ferreira França	Avaliação em larga escala: um estudo sobre erros dos alunos no trabalho com números e suas operações.	Dissertação (2008)
Alessandro Gonçalves	Análise das estratégias e erros dos alunos do 9º ano em questões de álgebra baseadas no Saresp de 2008 a 2011.	Dissertação (2014)
Cicero Santos e Paulo Oliveira	Avaliação externa em matemática: análise de teses e dissertações que abordam conteúdos matemáticos.	Artigo (2020)

Fonte: Elaborado pela autora (2022).

Objetivando organizar o arcabouço teórico utilizado nesta pesquisa, três linhas foram traçadas. A primeira delas diz respeito a um conjunto de estudos sobre avaliações em larga escala e seus impactos, que somaram sete estudos levantados. Em ordem cronológica, o primeiro deles é a dissertação de João Luiz Horta Neto, defendida em 2006, na qual o autor discorre sobre a "utilização da avaliação externa, especialmente o Saeb, como instrumento para a melhoria da qualidade educacional" (HORTA NETO, 2006, p. iv). A contribuição desse trabalho para a presente pesquisa se deu justamente na análise que faz sobre os principais resultados apresentados pelos relatórios nacionais dos resultados dessa avaliação, principalmente quanto à construção da escala de proficiência e sua interpretação.

Em 2010, Sandra Zákia Sousa e Romualdo Portela de Oliveira publicaram um artigo sobre sistemas estaduais de avaliação e como eles "vêm informando a formulação e implementação de políticas educacionais" (SOUSA; OLIVEIRA, 2010). Considerando que, segundo os autores, esses sistemas têm como principal referência o Saeb, a sua leitura foi importante para entender que, mesmo em iniciativas estaduais, apesar de teoricamente conseguirem se aproximar mais das

realidades vivenciadas pelas escolas, ainda são escassos os impactos das avaliações nas políticas de fomento ao ensino.

A autora Sandra Zákia Sousa aparece mais uma vez entre os estudos selecionados, dessa vez em coautoria com Alicia Bonamino, em artigo publicado em 2012 sobre as gerações de avaliação da educação básica no Brasil. Nesse trabalho, as autoras propõem três gerações, cujas diferenças residem no impacto de suas consequências, em termos de políticas de responsabilização para os agentes escolares. Se por um lado essa pesquisa critica o estreitamento curricular em razão de uma maior preparação dos estudantes para a realização de testes, por outro ela aponta o "potencial das avaliações de segunda e terceira gerações em propiciarem uma discussão informada sobre o currículo escolar, em termos das habilidades fundamentais de leitura e matemática que ainda não têm sido garantidas a todos os alunos" (BONAMINO; SOUSA, 2012, p. 373).

Também no ano de 2012, Adriana Bauer apresenta os resultados iniciais de um levantamento bibliográfico feito nas bases do Banco de Teses e Dissertações da Capes, com o objetivo de "analisar a produção acadêmica sobre avaliação de sistemas educacionais" (BAUER, 2012). Esse trabalho foi importante por oferecer uma visão global dessa temática nas pesquisas nacionais.

Horta Neto também contribui neste estudo com a sua tese, defendida em 2013, que teve como objetivo "analisar o desenvolvimento dos testes, aplicados pela União e pelos Estados de Minas Gerais e São Paulo [...], buscando identificar como os resultados obtidos estão sendo utilizados pelas políticas educacionais" (HORTA NETO, 2013). Em seu trabalho, cabe destacar, para os propósitos da presente pesquisa, o levantamento feito acerca das alterações que o Saeb sofreu ao longo dos anos e dos relatórios técnicos que divulgam os seus resultados.

Os estudos mais recentes que compõem essa linha de pesquisa foram publicados por Fátima Soares da Silva e Telma Ferraz Leal, em 2018, e por Rodrigo Marques, Ronildo Stieg e Wagner dos Santos, em 2020. O primeiro apresenta um estudo sobre a escala de proficiência da Prova Brasil, apontando inconsistências e problemas de progressão que dificultam a sua compreensão. O segundo traz um mapeamento de estudos sobre avaliação em larga escala e exames estandardizados em diferentes países, revelando como o tema tem sido explorado em pesquisas científicas internacionais.

Quanto à segunda linha de pesquisa, foram selecionados cinco estudos que, de uma maneira geral, abordam metodologias de construção de escalas de proficiência e de interpretação dessas escalas. O primeiro da lista, de Ruben Klein, publicado em 2003, trata sobre o modelo da TRI – de grupos múltiplos – adotado no Saeb "para a calibração dos itens e a para a obtenção de uma escala única, por disciplina, para as proficiências dos alunos das 4ª e 8ª séries do Ensino Fundamental e para a 3ª série do Ensino Médio, para os Saeb's a partir de 1995" (KLEIN, 2003). A partir da leitura desse estudo, é possível entender alguns dos mecanismos estatísticos utilizados em avaliações em larga escala

Do ano de 2007, dois artigos publicados na Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación trouxeram importantes contribuições para a presente pesquisa. O primeiro deles, de autoria de Lina Kátia Mesquita de Oliveira, Creso Franco e Tufi Machado Soares, apresenta uma nova proposta metodológica para a construção da escala de proficiência de matemática no âmbito do Estudo Longitudinal sobre a Qualidade e Equidade no Ensino Fundamental Brasileiro (Projeto GERES)[3], com o objetivo de superar limitações apontadas pelos autores como usuais nos programas de avaliação em larga escala. Nessa nova proposta,

> [...] são os itens que norteiam a interpretação da escala, e não os níveis de proficiência, a partir de uma análise detalhada das habilidades desenvolvidas pelos alunos, consideradas três fases importantes no processo de aprendizagem, quais sejam: (i) a introdução de uma habilidade / Início do desenvolvimento da habilidade; (ii) o processamento da habilidade / Rápido desenvolvimento da habilidade; e (iii) a consolidação da habilidade / Máximo desenvolvimento da habilidade (OLIVEIRA; FRANCO; SOARES, 2007).

Assim, ancoradas na metodologia usualmente utilizada em avaliações em larga escala, "as análises estatística e pedagógica conjugadas propiciaram a construção de uma escala de proficiência que apresenta as habilidades em início de desenvolvimento, em processamento ou em consolidação em um *continuum*" (OLIVEIRA; FRANCO; SOARES, 2007).

[3] O Projeto GERES/2005 consistiu em uma pesquisa na qual uma amostra de alunos e escolas de cinco cidades brasileiras foram observados ao longo de quatro anos. Participaram desse estudo cinco instituições de educação superior, a saber: Universidade Federal da Bahia (UFBA), Universidade Federal de Juiz de Fora (UFJF), Pontifícia Universidade Católica do Rio de Janeiro (PUC Rio), Universidade Federal de Minas Gerais (UFMG) e Universidade de Campinas (Unicamp).

O segundo artigo é de autoria de Nilma Santos Fontanive, Ligia Gomes Elliot e Ruben Klein e versa sobre os desafios da apresentação dos resultados de avaliações de sistemas escolares a diferentes públicos. Nesse trabalho, se objetivou o delineamento de estratégias que permitissem uma tradução do instrumental técnico e estatístico pouco familiar aos professores "em linguagens facilitadoras da sua compreensão, apoiadas em analogias, recursos visuais, ilustrações e exemplificações que concretizem os conceitos e aproximem a tecnologia de avaliação em larga escala do cotidiano, do próximo e do familiar ao contexto escolar" (FONTANIVE; ELLIOT; KLEIN, 2007). Para tanto, abordam os conceitos relacionados à escala de proficiência e à metodologia utilizada para sua interpretação.

A tese de Lina Katia Mesquita de Oliveira, defendida em 2008, traz um estudo sobre escalas de proficiências, investigando "diferentes abordagens para se obter uma associação entre os itens de um teste e os níveis ou pontos significativos de proficiência de diversas escalas atualmente empregadas", em nível nacional e internacional (OLIVEIRA, 2008). Ademais, propõe uma abordagem alternativa para caracterizar os níveis de uma escala, com base em métodos de análise de conglomerados[4], tendo como referência dois pontos notáveis da CCI, já abordados no seu artigo em 2007: o ponto onde a habilidade encontra-se no auge de seu desenvolvimento e o ponto onde se consolida a habilidade requerida para se acertar o item. Segundo a autora:

> Uma das vantagens decorrentes dessa metodologia reside no fato de que a interpretação da escala assim obtida é orientada pelo agrupamento dos itens com base em suas respectivas curvas características, e não pela seleção de itens a partir de níveis de proficiência predeterminados. Tal método tinha, entre outros inconvenientes, o fato de apresentar uma grande incerteza prévia quanto ao número de itens considerados para cada nível de referência (OLIVEIRA, 2008, p. 143).

Por fim, o último estudo selecionado nessa linha é a dissertação da Flávia Almeida Perry, defendida em 2009, que teve como objetivo "apresentar e discutir métodos de construção de escalas de proficiência e as abordagens de interpretação das escalas na área de língua portuguesa" (PERRY, 2009). Para tanto, a autora utilizou a metodologia empregada no Saeb, no Projeto GERES e o método de *cluster*. As três escalas construídas foram, então, apresentadas a professores

[4] A análise de agrupamento ou *clusters* é um conjunto de técnicas multivariadas de interdependência, cuja finalidade primária é agregar elementos segundo o grau de semelhança entre si e segundo a maior dessemelhança possível em relação aos objetos pertencentes a outros grupos (OLIVEIRA, 2008).

para que eles dessem sua opinião acerca do conteúdo e da forma como eram apresentadas. Essa análise revelou uma lacuna na formação de professores, seja ela inicial ou continuada, em relação aos processos envolvidos em avaliações em larga escala.

Como referencial teórico para a terceira linha de pesquisa, foram investigados estudos disponíveis no Google Acadêmico utilizando as palavras-chaves erro e avaliação em larga escala. Entre os achados, que pertenciam majoritariamente à área da matemática, foram selecionados três trabalhos, os quais guardam estreita relação com a análise de erros em avaliações em larga escala. Ressalta-se que não foram encontrados estudos sobre análise de erros em língua portuguesa, componente curricular que é foco da presente dissertação.

O primeiro deles, de autoria de Maria José Ferreira França, consiste em uma dissertação sobre os erros dos estudantes no trabalho com números e suas operações em avaliação em larga escala, de 2008. Nesse trabalho, a autora buscou analisar a natureza dos erros cometidos por estudantes da 4ª série do ensino fundamental do Estado de Pernambuco, a partir dos percentuais de marcação em cada um dos itens selecionados para a pesquisa.

O segundo, de Alessandro Gonçalves, faz uma análise das estratégias e erros dos estudantes do 9º ano em itens de álgebra baseados no Sistema de Avaliação da Aprendizagem Escolar do Estado de São Paulo (Saresp) de 2008 a 2011. Nesse caso, o pesquisador selecionou itens da avaliação e os reaplicou com um pequeno grupo de estudantes no contexto da sala de aula, a fim de investigar as estratégias adotadas pelos estudantes na resolução dos problemas.

Finalmente, o terceiro trabalho, publicado na Revista Brasileira de Iniciação Científica, traz o estado da arte de teses e dissertações que abordam a avaliação externa e conteúdos matemáticos no período de 2001 a 2017. Em sua pesquisa, Cicero Santos e Paulo Oliveira identificaram que os autores pesquisados “analisaram as estratégias, dificuldades e erros de alunos ao lidar com conceitos” (SANTOS; OLIVEIRA, 2020), servindo de grande ajuda para entender o panorama da pesquisa sobre erros em avaliações em larga escala, objeto do presente estudo.

A partir desse arcabouço teórico, o presente estudo propõe uma pesquisa de natureza quali-quanti, esquematizada no Quadro 2 a seguir.

Quadro 2 - Quadro de coerência da pesquisa

Tema da pesquisa		
O erro como ferramenta para compreender o desempenho de estudantes em avaliações em larga escala		
Questão geral		
É possível utilizar as marcações dos estudantes nas opções erradas para gerar evidências sobre o seu desempenho em avaliações em larga escala?		
Objetivo geral		
Analisar os erros cometidos pelos estudantes em um teste de avaliação em larga escala		
Questões secundárias	**Objetivos específicos**	**Metodologia**
Como se desenvolveu a construção de escalas de proficiência no Brasil?	Verificar de que maneira têm sido concebidas e conduzidas as escalas de proficiência do Saeb.	Análise documental e bibliográfica
Como construir uma escala interpretada que leve em consideração o erro do estudante em avaliações em larga escala?	• Construir uma escala que leve em consideração o erro do estudante na aplicação de um teste de língua portuguesa. • Gerar evidências de intervenção pedagógica a partir da interpretação dessa escala.	Análises estatísticas com base nos modelos nominal e de respostas graduais da TRI e análise de conteúdo dos itens com base na matriz de referência da avaliação

Fonte: Elaborado pela autora (2022).

Como se pode depreender do quadro, **o objetivo geral da presente dissertação é analisar os erros cometidos pelos estudantes em um teste de avaliação em larga escala**. Para tanto, esta dissertação foi organizada em dois estudos independentes e complementares.

No estudo I, objetiva-se **verificar de que maneira têm sido concebidas e conduzidas as escalas de proficiência do Saeb**, a partir da análise de documentos e relatórios técnicos publicados pelo Instituto Nacional de Pesquisas Educacionais Anísio Teixeira (Inep), respondendo à pergunta *como se desenvolveu a construção de escalas de proficiência no Brasil?* A escolha pela escala de proficiência do Saeb se deu em vista dessa ser a principal avaliação em larga escala conduzida no Brasil, que serve de referência para diversas outras iniciativas de avaliação nos estados e municípios do país. Espera-se, com este trabalho, entender as metodologias que permitem a criação da escala de proficiência do Saeb e como essa escala se desenvolveu ao longo dos ciclos de aplicação do teste.

Já no estudo II, pretende-se **construir uma escala interpretada que leve em consideração o erro do estudante na aplicação de um teste de língua portuguesa**, invertendo

a lógica das tradicionais escalas de proficiência. Retomando metodologias utilizadas nacional e internacionalmente, o objetivo é chegar a uma proposta que viabilize a construção dessa escala, respeitando critérios de validade e confiabilidade. Dessa forma, ao invés de mostrar as habilidades que os estudantes de determinado nível provavelmente desenvolveram, se pretende apontar as deficiências que ainda caracterizam o seu desempenho. Assim, espera-se responder à pergunta *como construir uma escala interpretada que leve em consideração o erro do estudante em avaliações em larga escala?*

Para alcançar esse objetivo, foram utilizados dados provenientes da Prova São Paulo[5], uma iniciativa de avaliação em larga escala que, apesar de ser municipal, se espelha nos moldes do Saeb para medir o desempenho dos seus estudantes em língua portuguesa e matemática. Reforça-se que a opção pelo uso desses dados se deve ao fato de que, mediante solicitação enviada à Secretaria de Municipal de Ensino de São Paulo (Anexo I), foi dado acesso a todos os dados necessários para a condução da presente pesquisa, sejam eles abertos ou sigilosos.

A intenção, com isso, é mapear os erros usualmente cometidos por esse grupo de estudantes, caracterizando os distintos perfis de desempenho na referida avaliação. A partir dessa escala, seria possível delinear, por exemplo, quais são os erros que, em geral, estudantes com determinada proficiência cometem. Em outras palavras, se até então o professor recebia a informação de quais habilidades seus estudantes demonstraram possuir, com essa escala seria possível indicar os erros que cometem no processo de desenvolvimento das habilidades que ainda não alcançaram. **Essa escala interpretada corresponde ao produto técnico resultante da presente pesquisa**, que tem como objetivo oferecer à comunidade escolar e à sociedade uma devolutiva a respeito do desempenho dos estudantes em avaliações em larga escala, auxiliando-os a compreender os resultados dessas avaliações e utilizá-los para melhorar o desempenho dos estudantes e, consequentemente, a qualidade da educação.

Com este trabalho, se espera desenvolver formas alternativas de compreender os resultados de avaliações em larga escala e dirimir as dificuldades enfrentadas por muitos professores da rede pública em utilizar os seus resultados no planejamento escolar. Advém dessa missão e da busca

[5] A Prova São Paulo é uma avaliação censitária, destinada aos estudantes do 4º ao 9º ano do ensino fundamental da rede municipal de ensino de São Paulo. É avaliado o desempenho em língua portuguesa (leitura e produção de textos), matemática e ciências naturais, tendo como base as Matrizes de Referência da Avaliação do Rendimento Escolar da Rede Municipal de Ensino de São Paulo.

por um maior aprofundamento nas análises pedagógicas, englobando tanto o acerto quanto o erro dos estudantes, a justificativa para esta pesquisa.

ESTUDO I

AS METODOLOGIAS DE CONSTRUÇÃO DAS ESCALAS DE PROFICIÊNCIA DO SAEB DE 1995 A 2017

Introdução

Na educação básica brasileira, a avaliação foi instituída como política de Estado e estratégia de monitoramento do ensino em meados de 1990, com a criação do Sistema de Avaliação da Educação Básica (Saeb) pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), vinculado ao Ministério da Educação (MEC). Desde então, e a cada dois anos, são aplicados testes aos estudantes dos ensinos fundamental e médio, com vistas a realizar um diagnóstico da educação no país. Em 1996, essa política foi reforçada pela Lei de Diretrizes e Bases da Educação Básica (Lei n. 9.394/1996), que incumbiu à União o papel de "assegurar processo nacional de avaliação do rendimento escolar no ensino fundamental, médio e superior, em colaboração com os sistemas de ensino, objetivando a definição de prioridades e a melhoria da qualidade do ensino" (BRASIL, 1996).

Apesar de terem sido feitas tentativas de testar outras áreas de conhecimentos, o Saeb avalia primordialmente as áreas de língua portuguesa (com foco em leitura) e de matemática (com foco na resolução de problemas). São avaliados os últimos anos de cada etapa de ensino – o 5º e o 9º ano do ensino fundamental e o 3º ano do ensino médio.

Para cada uma dessas etapas, além de questionários contextuais que investigam fatores intra e extraescolares que influenciam o desempenho, são aplicados testes cognitivos para medir a proficiência dos estudantes da rede pública (censitariamente) e da rede privada (amostralmente). A partir desses dados quantitativos, a metodologia utilizada nas análises das respostas dos indivíduos a esses testes – a Teoria de Resposta ao Item (TRI) – permite a construção da escala de proficiência, um documento com teor qualitativo, que visa associar a valores numéricos um conjunto de informações pedagógicas.

A escala de proficiência contém uma escala numérica que representa as proficiências dos estudantes em um *continuum*, divididas em intervalos ou níveis, e um conjunto de descrições que visam traduzir essas proficiências em termos de habilidades ou competências que os estudantes demonstraram ao realizar o teste (FONTANIVE; ELLIOT; KLEIN, 2007).

Esse é, portanto, um resultado que, para além das médias e percentuais – dados quantitativos –, oferece uma interpretação pedagógica do desempenho dos indivíduos – dado qualitativo. Por isso, era de se esperar que ele fosse amplamente divulgado e utilizado pela comunidade escolar, principalmente professores e gestores. No entanto, diversos estudos revelam que, por mais que tenham algum conhecimento dessa escala, são poucos os que conseguem compreendê-la e efetivamente utilizá-la no planejamento didático (PERRY, 2009; SOUSA; OLIVEIRA, 2010; KISTEMANN JR; GOUVÊA, 2019; CALDERÓN; BORGES, 2020). Por outro lado, há também estudos que apontam incoerências e deficiências no processo de elaboração das escalas que justificariam essas dificuldades (HORTA NETO, 2006; SILVA; LEAL, 2018).

À parte das discussões sobre o uso de resultados de avaliações em larga escala ou das metodologias adotadas, parece ser consenso nos universos acadêmico e governamental que uma quantidade expressiva de dados tem sido coletada em avaliações nacionais, estaduais e municipais, porém, poucos estudos têm se dedicado a investigar seus resultados e impactos (FONTANIVE; ELLIOT; KLEIN, 2007; SOUSA; OLIVEIRA, 2010).

Com o intuito de contribuir nesse cenário, o presente estudo pretende trazer uma sistematização das escalas de proficiência do Saeb, principal avaliação da educação básica no Brasil, no intuito de esclarecer o processo de construção desses documentos e as mudanças pelas quais passaram desde a sua criação, em 1995. **O objetivo geral é verificar de que maneira têm sido concebidas e conduzidas as escalas de proficiência do Saeb**. Já os objetivos específicos são: (i) investigar as metodologias que embasam a construção das escalas de proficiência do Saeb; (ii) analisar comparativamente a metodologia de construção das escalas de proficiência publicadas pelo Inep e relacionadas ao Saeb; e (iii) identificar aspectos positivos e negativos no processo de evolução das escalas de proficiência do Saeb.

Método

Para cumprir com esse propósito, o presente estudo apresenta uma pesquisa bibliográfica e documental com abordagem qualitativa, que visa aprofundar o conhecimento a respeito do objeto em foco, a saber, as escalas de proficiência do Saeb construídas e interpretadas. A partir dessa análise bibliográfica, se buscou compreender a teoria que embasou o processo de construção dessas escalas e sua interpretação pedagógica. Para tanto, se utilizaram referências nacionais e internacionais na área de avaliação educacional, como os trabalhos publicados por Hambleton, Swaminathan e Rogers (1991), Beaton e Allen (1992), Klein e Fontanive (1995), Valle (2000), Andrade, Tavares e Valle (2000), Pasquali e Primi (2003), Horta Neto (2006), Oliveira (2008), Araujo, Andrade e Bortolotti (2009), entre outros. Já na análise documental, investigaram-se os relatórios técnicos publicados pelo Inep a respeito do Saeb, além de documentos oficiais pertinentes à avaliação educacional no país. O objetivo é fazer uma análise comparativa das metodologias de construção das escalas de proficiência da principal avaliação em larga escala realizada no Brasil.

Como resultados, primeiramente, são apresentados os pressupostos que fundamentam as escalas de proficiências, abordando a metodologia que viabiliza a sua construção. Em seguida, são analisadas as escalas de proficiência do Saeb no período entre 1995 e 2017, dando ênfase aos aspectos técnicos e formais de sua constituição. Por fim, se apresentam as conclusões deste estudo.

Resultados e discussão

1. Aspectos metodológicos da construção de escalas de proficiência

Começando pelos pressupostos que fundamentam a criação de escalas de proficiência, há que se mencionar a metodologia utilizada para viabilizar esse procedimento, a Teoria de Resposta ao Item (TRI). Essa teoria surgiu da necessidade de superar dificuldades enfrentadas na análise das respostas dos indivíduos a um teste pela Teoria Clássica dos Testes (TCT), limitada no quesito comparabilidade tanto entre os próprios indivíduos quanto entre diferentes aplicações de um teste. Isso porque, na TCT,

> [...] os resultados encontrados dependem do particular conjunto de itens que compõem o instrumento de medida, ou seja, as análises e interpretações estão sempre associadas à prova como um todo. Assim, torna-se inviável a comparação entre indivíduos que não foram submetidos às mesmas provas, ou pelo menos, ao que se denomina de formas paralelas de testes (VALLE, 2000, p. 7).

Da mesma maneira, os resultados alcançados na TCT em determinado teste dependem da população que participou, uma vez que mais difícil será o teste se menos proficiência tiverem os sujeitos que responderam e mais fácil ele será se maior for essa proficiência (PASQUALI; PRIMI, 2003). Sendo assim, a análise pela TCT está atrelada ao contexto de aplicação do teste, o que inviabiliza, por exemplo, a construção de uma série histórica que permita observar a evolução no desempenho ao longo dos anos ou até mesmo a definição do nível de dificuldade do item, independentemente da população que responde a esse teste.

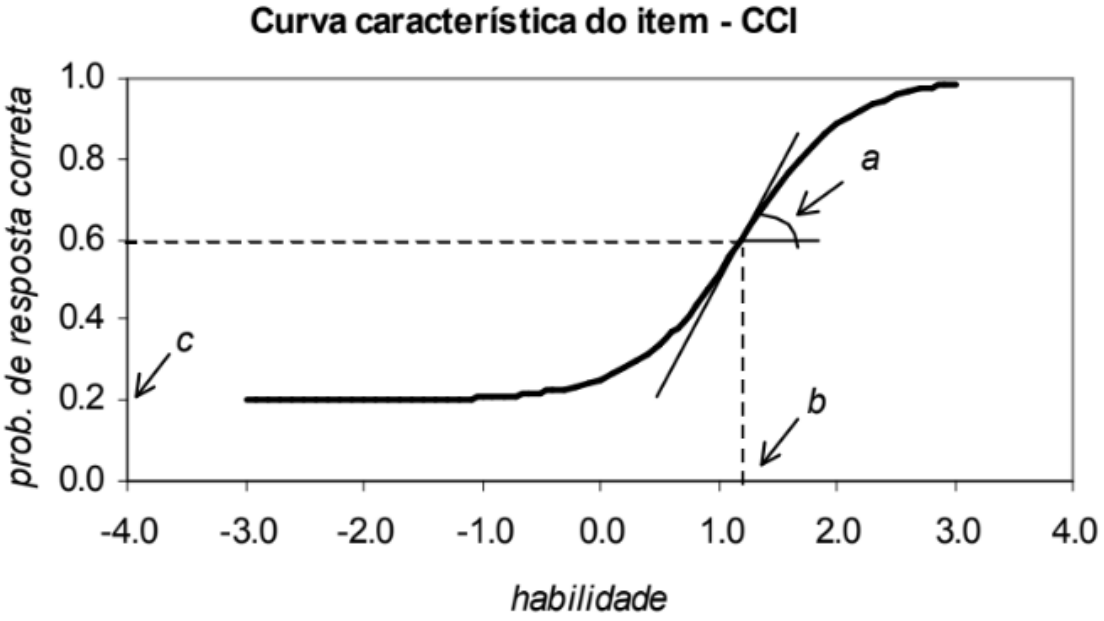
Ainda na década de 50, deu-se início à elaboração de um modelo teórico e métodos para estimar parâmetros em testes que superassem essas limitações. Na década de 80, esses métodos alcançaram o seu auge, em razão do desenvolvimento de tecnologias e *softwares* que tornaram possíveis os cálculos matemáticos envolvidos na aplicação da TRI (PASQUALI; PRIMI, 2003).

A TRI abrange uma série de modelos que visam explicar, probabilisticamente, o desempenho de indivíduos perante um determinado teste. Consiste, portanto, em um conjunto de modelos matemáticos que utiliza o item como unidade básica de análise. Com a TRI, se torna possível comparar indivíduos que realizam o mesmo teste em determinado ano ou em anos diferentes, estimar a proficiência dos indivíduos com base em outros parâmetros que não só a quantidade de acertos e calcular os parâmetros independentemente dos sujeitos que respondem o teste (ARAUJO; ANDRADE; BORTOLOTTI, 2009).

Apesar dos seus diversos modelos, em avaliação educacional é comumente utilizado o modelo logístico da TRI de três parâmetros, que calcula o nível de dificuldade dos itens do teste, o poder de discriminação de cada um deles em relação à habilidade investigada e a probabilidade de acerto ao acaso do indivíduo. Esses parâmetros são estatisticamente denominados *a*, *b* e *c* e a função que calcula a probabilidade de um indivíduo responder corretamente um item em função deles e da sua habilidade é representada pela Curva Característica do Item (CCI), ilustrada a título de exemplificação na Figura 1, a seguir. De modo geral, se observa que à medida que a proficiência

do indivíduo aumenta, também aumenta a probabilidade de ele responder o item corretamente (ANDRADE; TAVARES; VALLE, 2000).

Figura 1 - Exemplo de Curva Característica do Item



Fonte: Andrade, Tavares e Valle (2000).

Em se tratando especificamente do que cada um desses parâmetros evidencia a respeito dos itens de um teste, o parâmetro *a* representa a discriminação do item, indicando o quanto ele é capaz de diferenciar indivíduos que possuem a habilidade avaliada daqueles que não a possuem (HAMBLETON; SWAMINATHAN; ROGERS, 1991). Um valor maior que zero indica uma discriminação positiva, ou seja, que é provável que um indivíduo que acerta determinado item de fato possua a habilidade avaliada. Quanto mais alto for esse valor, maior é o poder de discriminação do item.

O parâmetro *b* diz respeito à dificuldade do item, normalmente expressa em uma escala de -3 a 3. Quanto maior for o valor de b, maior é a habilidade requerida para alcançar essa probabilidade de acerto – isso significa que mais difícil é o item. Em contrapartida, quanto menor

for o valor de b, menor é a habilidade e, consequentemente, mais fácil é o item. (HAMBLETON; SWAMINATHAN; ROGERS, 1991).

Por fim, o parâmetro *c* calcula a chance de um indivíduo acertar um item mesmo que ele não possua a habilidade que está sendo avaliada (HAMBLETON; SWAMINATHAN; ROGERS, 1991). Essa medida visa identificar, por exemplo, itens que, por problemas em sua construção, talvez estejam dando pistas sobre a resposta correta, de modo que o indivíduo pode acertar, apesar de não dominar aquele conteúdo, provavelmente, utilizando estratégias outras que não o seu próprio conhecimento.

A partir do cálculo desses parâmetros para todos os itens, realizados por *softwares* específicos e adotando técnicas estatísticas e matemáticas de estimação (ARAUJO; ANDRADE; BORTOLOTTI, 2009), se identificam aqueles que apresentam problemas. Por exemplo, o item pode apresentar baixa discriminação, não revelando se os indivíduos que o acertaram de fato possuem aquela habilidade, ou pode apresentar alta probabilidade de acerto ao acaso, mostrando que os indivíduos que acertaram o item podem tê-lo feito por mero acaso ou "chute". Esses itens normalmente são excluídos das análises. Destaca-se que essa exclusão visa garantir a validade do cálculo como um todo, retirando do *corpus* as informações que possam enviesar os resultados. Considerando que se trata de uma análise probabilística, tais itens podem ser considerados como "margem de erro", que é inerente a qualquer medida que envolva estimativas.

Após esse filtro, se estimam as proficiências a partir dos erros e acertos em cada um dos itens aplicados e do padrão de resposta de cada indivíduo. Nesse momento, tanto os parâmetros dos itens quanto as proficiências dos indivíduos são colocados em uma mesma escala, que tem como referência os valores de dificuldade (parâmetros *b*) dos itens. Assim, torna-se possível associar, por meio dos itens, as proficiências calculadas e as habilidades dos indivíduos (ANDRADE; TAVARES; VALLE, 2000).

A metodologia para a construção dessa escala foi descrita por Beaton e Allen (1992), que definem os critérios para a seleção do que denominam níveis âncora. Segundo os autores, esses níveis devem ser previamente definidos de acordo com os objetivos e as propriedades da escala. Cada um deles é constituído por um conjunto de itens que atendem três requisitos: (i) pelo menos 65% dos indivíduos de determinado nível acertam o item; (ii) no máximo 50% dos indivíduos no

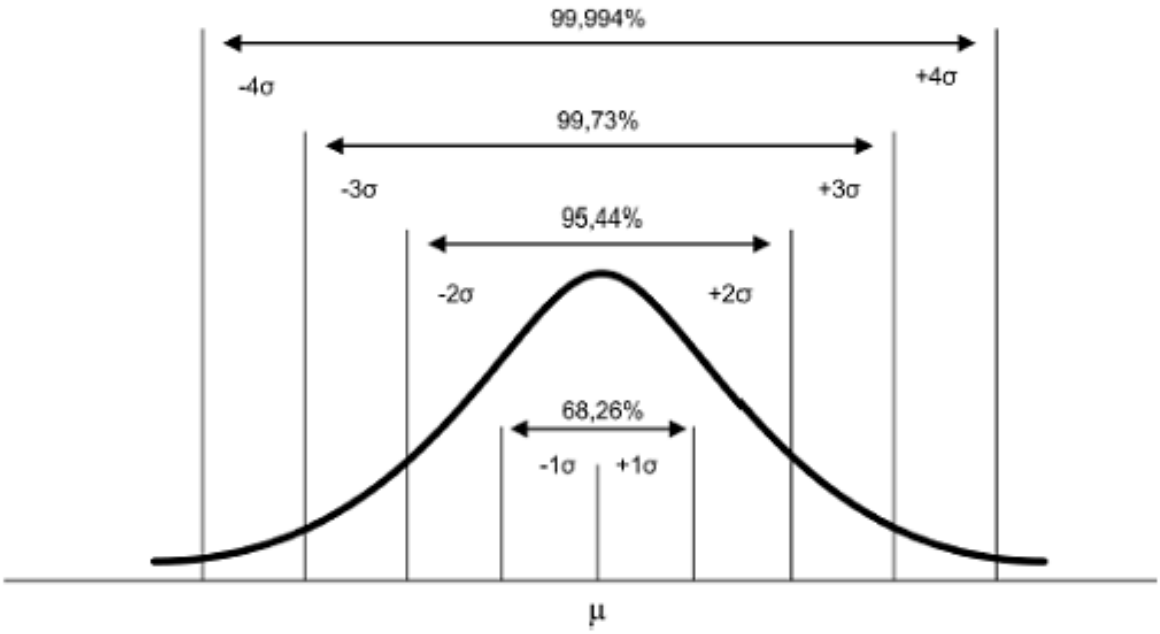
nível imediatamente anterior acertam o item; e (iii) a diferença entre esses dois percentuais é de pelo menos 30% (BEATON; ALLEN, 1992).

Essa escala, que inicialmente é construída no intervalo de -3 a 3 (conforme os valores de *b* dos itens), geralmente passa por uma transformação linear, que tem objetivo puramente didático, com vistas a facilitar a interpretação dos dados ao fornecer resultados com valores positivos (ANDRADE; TAVARES; VALLE, 2000).

A distribuição esperada dos itens e dos indivíduos nessa escala se assemelha a uma curva de Gauss[6], que prevê cerca de 68% concentrados a um desvio padrão da média, tanto para cima quanto para baixo. Dentro de dois desvios padrões para cima e para baixo, estariam aproximadamente 95% dos indivíduos e dos itens, e no espaço de 3 desvios padrões estaria quase a totalidade da população alvo e dos itens aplicados. Essa distribuição pode ser mais bem visualizada na Figura 2, em que **μ** representa a média e **σ**, o desvio padrão.

[6]A curva de distribuição normal, também chamada de curva de Gauss, consiste em um gráfico de densidade utilizado como modelo para representar o comportamento de fenômenos aleatórios com base em dois parâmetros: média e desvio-padrão (SALKIND, 2007).

Figura 2 - Representação da curva normal de Gauss



Fonte: Wikipedia (2016).

Cabe destacar que, para a construção de uma escala de proficiência, é importante que muitos itens tenham sido aplicados, de modo a garantir maior representatividade das habilidades avaliadas, e que os níveis escolhidos não sejam muito próximos (BEATON; ALLEN, 1992; ANDRADE; TAVARES; VALLE, 2000). Ressalta-se que, ainda que esses critérios sejam os ideais, observa-se muitas vezes que os níveis extremos da escala, referentes às proficiências mais baixas e às mais altas, são, de algum modo, mal representados, pois são constituídos por itens muito fáceis ou muito difíceis e ambos, em geral, são difíceis de elaborar, por isso, pouco frequentes em bancos de itens.

O próximo passo na construção de uma escala de proficiência é a sua interpretação pedagógica. A necessidade dessa etapa advém do fato de que se deter apenas na informação da localização dos respondentes em determinada escala traz respostas quantitativas, mas nenhuma informação qualitativa. Por exemplo, observando dois indivíduos que estejam nas posições 200 e 275 de uma escala com média 250 e desvio padrão 25, é possível concluir que aquele localizado no ponto 200 está 2,0 desvios padrão abaixo da média e que o que está localizado em 275 está 1,0

desvio-padrão acima da média. Nessa situação, provavelmente, o segundo indivíduo demonstrou possuir mais habilidades do que o primeiro (ANDRADE; TAVARES; VALLE, 2000).

No entanto, para melhor compreender essa diferença, é importante introduzir informações qualitativas, que tornem possível a utilização didática dos dados quantitativos da escala. A interpretação pedagógica de uma escala normalmente é feita por um grupo de especialistas da temática avaliada, que, de posse dos itens que compuseram o teste ou conjunto de testes, bem como dos resultados a eles associados, identificam as habilidades necessárias para responder corretamente cada item, conforme a matriz de referência da avaliação (BEATON; ALLEN, 1992).

É importante destacar que, por matriz de referência, se entende o documento que norteia uma avaliação educacional. De maneira geral, consiste em um recorte do que é mais relevante e possível de ser avaliado em um teste em larga escala no currículo de determinada área do conhecimento, delimitando o que se espera que os estudantes dominem ao longo de determinado período escolar (HORTA NETO, 2006).

Com base na matriz, ao final da interpretação de uma escala de proficiência, se espera ter um detalhamento do desempenho dos indivíduos, indicando aquilo que, por meio do teste aplicado, demonstraram ter desenvolvido em termos de habilidade. Vale ressaltar que esses dados, desde os parâmetros dos itens até a descrição da escala, são gerados com base em cálculos probabilísticos (ANDRADE; TAVARES; VALLE, 2000). Sendo assim, se admite o seu teor aproximativo, uma vez que eles nada revelam além de uma probabilidade em relação às habilidades avaliadas no teste. Isso quer dizer que a interpretação dos resultados deve levar em conta que as descrições de determinado nível de desempenho, provavelmente, representam as habilidades que os indivíduos daquele nível possuem. Essa representação constitui apenas parte do conhecimento desse indivíduo, justamente aquela que se pretendeu avaliar no teste (HORTA NETO, 2013).

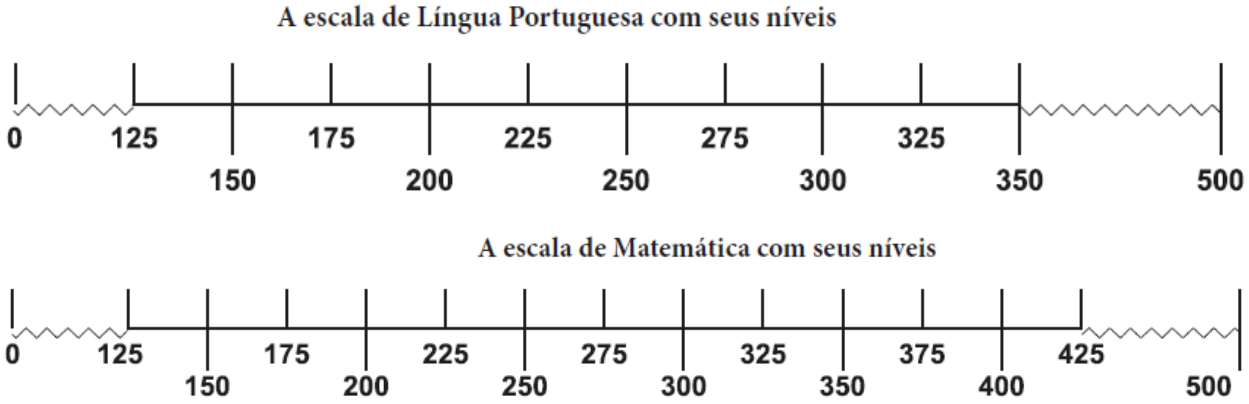
Ademais, cabe acrescentar que as descrições de cada nível são cumulativas, isto é, aqueles indivíduos que foram posicionados nos níveis superiores da escala, provavelmente, possuem as habilidades descritas naquele nível e as descritas em todos os níveis anteriores (BEATON; ALLEN, 1992).

2. Análise das escalas de proficiência do Saeb (1995-2017)

Após a discussão dos métodos que embasam a construção de escalas de proficiência, se apresenta a análise das escalas de proficiência do Saeb e de que maneira evoluíram ao longo do tempo. A seguir, serão explicitados os critérios adotados em cada ciclo do Saeb para a definição dos pontos-âncora, a caracterização de um item também como âncora e, por fim, a interpretação da escala de proficiência. Espera-se, com isso, montar um histórico que esclareça os caminhos percorridos e as decisões tomadas ao longo desse processo.

Em 1995, com a adoção de novas técnicas e metodologias com base na TRI para a análise das respostas dos estudantes que participaram do Saeb, se deu início à avaliação dos indivíduos por meio da proficiência demonstrada por eles durante o teste (HORTA NETO, 2006). Para tanto, foi definida uma escala de proficiência que variava entre 0 e 500 pontos, ilustrada na Figura 3 a seguir.

Figura 3 - Representação da escala Saeb de proficiência em língua portuguesa e matemática

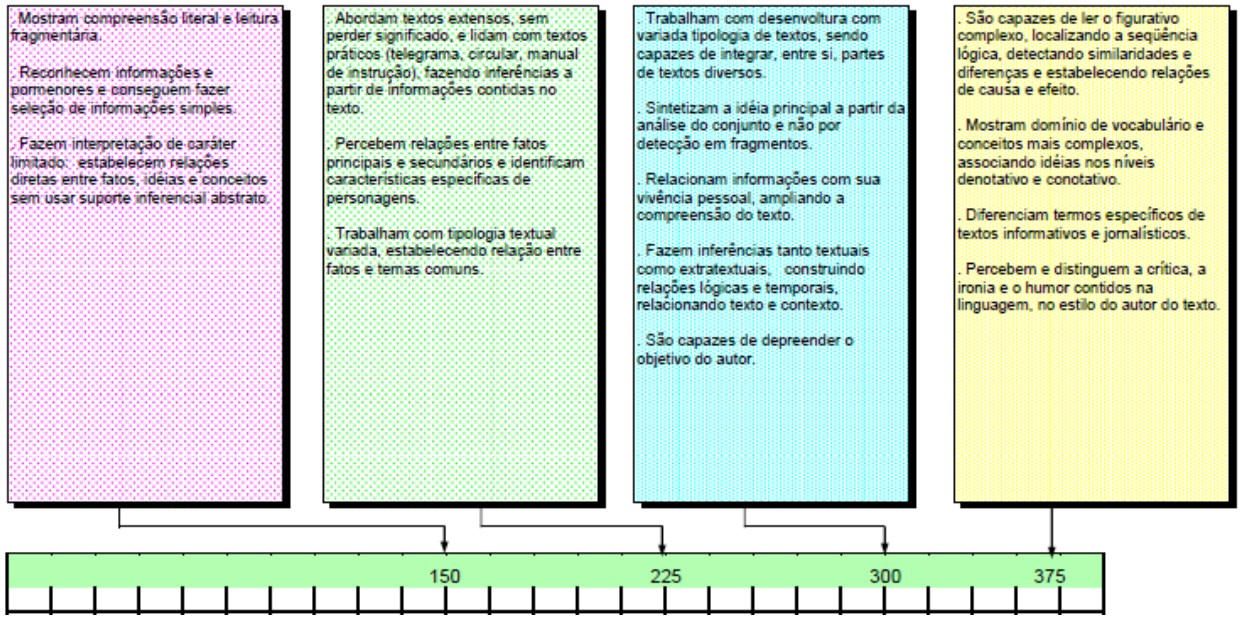


Fonte: Inep (2006).

Na inauguração da escala de proficiência do Saeb, quatro pontos foram definidos para a interpretação: 150, 225, 300 e 375. Um item era considerado representativo de um desses pontos se atendesse aos seguintes critérios: (i) um percentual de 65% ou mais de acerto no ponto sob análise; (ii) um percentual de acerto do item no ponto imediatamente inferior ao sob análise de até 25%; e (iii) um percentual de acerto do item no ponto imediatamente superior ao sob análise de pelo menos 95% (INEP, 1998).

Depois de posicionados nesses pontos, os itens foram submetidos a uma banca de especialistas, que, de posse do seu conhecimento técnico em cada área, os descreveram pedagogicamente conforme os seus respectivos pontos, indicando, para cada um deles, o que os indivíduos demonstraram saber, compreender ou ser capaz de fazer (INEP, 1998). Visualmente, a escala interpretada foi publicada em formato de quadros e cada ponto abarcou determinada quantidade de descrições, inseridas por meio de marcador textual e iniciadas por verbo na 3ª pessoa do plural no presente, conforme apresentado na Figura 4.

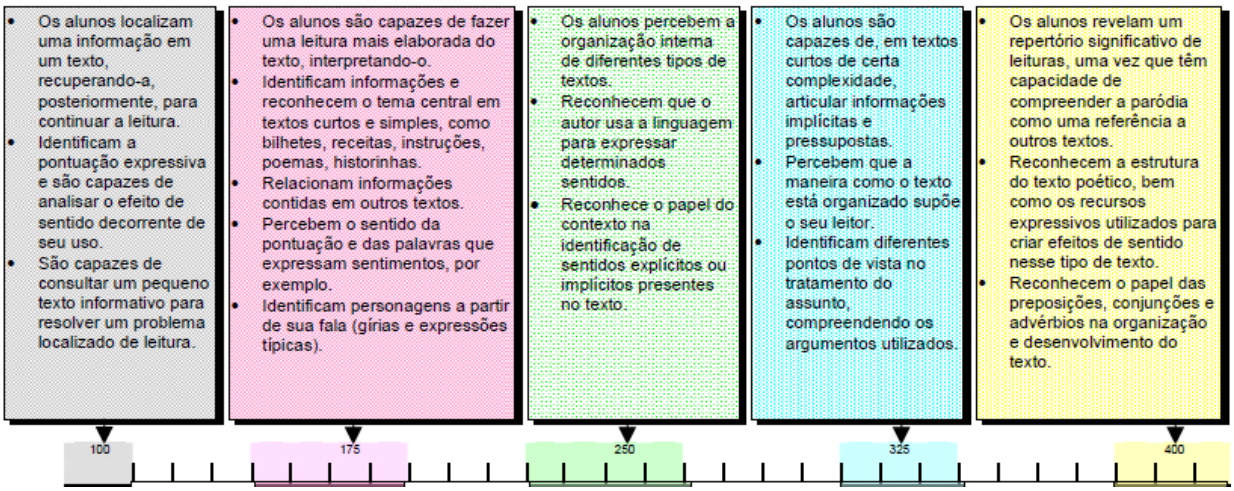
Figura 4 - Escala de proficiência de língua portuguesa do ciclo Saeb 1995



Fonte: Inep (1998)

No ciclo do Saeb 1997, houve mudança no método adotado para a ancoragem dos itens. Nesse ano, um item-âncora foi definido como sendo aquele que (i) 65% ou mais dos indivíduos em determinado ponto o acertam; (ii) menos de 50% dos indivíduos posicionados no ponto anterior o acertam e (iii) a diferença entre esses percentuais é maior que 30% (INEP, 1999). Além dessa alteração, os pontos escolhidos para a interpretação da escala foram aumentados em 25 pontos e passaram a ser 175, 250, 325 e 375. Para língua portuguesa, foi ainda inserido o ponto 100. Com os itens posicionados, o procedimento para a interpretação foi basicamente o mesmo, com a análise dos itens por um painel de especialistas que determinaram os conhecimentos e habilidades que os indivíduos demonstraram possuir quando situados em torno dos pontos estabelecidos (INEP, 1999). A escala interpretada foi igualmente organizada em formato de quadros, com o uso de marcadores textuais e verbos na 3ª pessoa do plural no presente, porém, com a inclusão das palavras “Os alunos” no primeiro marcador de cada nível, como se pode observar pela Figura 5 a seguir.

Figura 5 - Escala de proficiência de língua portuguesa do ciclo Saeb 1997



Fonte: Inep (1999)

Especificamente nesse ciclo, foi feita uma tentativa no sentido de associar os pontos da escala às etapas da educação básica. O ponto 250, por exemplo, corresponderia ao final do 2º ciclo do ensino fundamental em língua portuguesa e ao final do 1º ciclo do ensino fundamental em matemática. Essas informações podem ser visualizadas no Quadro 3 a seguir.

Quadro 3 - Relação entre os pontos da escala e os ciclos dos níveis de ensino proposta no relatório técnico do Saeb 1997

Nível de proficiência – escala Saeb/97	**Ciclo e nível de ensino**		
	Matemática	**Língua portuguesa**	**Ciências (Física, Química e Biologia)**
100	Não significativo	Até a metade do 1º ciclo do EF	Até a metade do 1º ciclo do EF
175	Até a metade do 1º ciclo do EF	Até o final do 1º ciclo do EF	Até o final do 1º ciclo do EF
250	Até o final do 1º ciclo do EF	Até o final do 2º ciclo do EF	Até a metade do 2º ciclo do EF
325	Até o final do 2º ciclo do EF	Até o final do EM	Até o final do 2º ciclo do EF
400	Até o final do EM	Além do final do EM	Até o final do EM

Fonte: Inep *apud* Horta Neto (2006).

Em razão disso, se encontra ainda no relatório publicado pelo Inep uma série de textos que se somam à leitura da interpretação da escala no intuito de esclarecer o significado pedagógico de cada um dos seus pontos em relação aos níveis de ensino. Por exemplo, quanto ao ponto 175 em língua portuguesa, esse relatório informa que

> [...] esse nível corresponde, em termos curriculares, ao que é proposto até o final do 1º ciclo do ensino fundamental (4ª série). Os dados mostram que menos da metade dos alunos da 4ª série são capazes de, por exemplo, relacionar informações dadas em um texto com experiências pessoais e com informações contidas em outros textos. Era de se esperar que houvesse um número maior de alunos da 4ª série com desempenho superior ao descrito nesse nível; na 8ª série do ensino fundamental, podem ser considerados satisfatórios os percentuais de alunos que superam esse nível (entre 89% e 95% dos alunos), sendo capazes, portanto, de fazer uma leitura direta de textos curtos, retirando informações e identificando o tema. Apesar da defasagem existente entre o desempenho real e o esperado pelos currículos, esses dados demonstram a consolidação, na 8ª série, da aprendizagem básica preconizada pelos currículos para os alunos da 4ª série do ensino fundamental (INEP, 1999, p. 15).

Segundo Horta Neto (2006, p. 73), a explicação desse procedimento no relatório técnico do Saeb 1997 se resumiu à indicação de que "foi feito primeiramente um cruzamento entre a matriz de referência e os resultados dos alunos, sendo a seguir consultados diversos especialistas nas áreas avaliadas". Esse mesmo autor apresenta argumentos contrários a essa iniciativa, pois

> [...] em um novo ciclo de avaliação a descrição das habilidades associadas a cada ponto não são necessariamente as mesmas daquelas descritas no ciclo anterior, já que as questões que estarão associadas a esses pontos [...] podem ser diferentes. Com isso, pode ser que não seja mantida a mesma relação entre o ponto da escala e o momento da aprendizagem existente no ciclo anterior. Outra razão para a inadequação [...] advém do fato de que existe uma grande dispersão nos resultados da avaliação. Com isso, é possível encontrar alunos de uma mesma série com diversas proficiências, e assim sendo, fica muito difícil associá-las às etapas da educação básica (HORTA NETO, 2006, p. 73).

Ainda conforme Horta Neto (2006, p. 73), "provavelmente pelas dificuldades aqui apontadas, no próximo ciclo de avaliação é abandonada a tentativa de associar os pontos da escala com as etapas da educação básica". Um outro comentário a respeito da escala de proficiência do ciclo 1997 do Saeb diz respeito à impossibilidade de comparação com a escala do ciclo de 1995, haja vista a mudança nos pontos escolhidos para a interpretação.

Em 1999, houve novamente uma mudança de critérios para a ancoragem de itens, em virtude da constatação de que os anteriores eram muito rigorosos e culminavam na exclusão de muitos deles, restando poucos para a interpretação da escala (OLIVEIRA, 2008). Quanto ao método de ancoragem desse ano, um item era alocado em determinado ponto, se: (i) naquele ponto houvesse mais de 50 indivíduos; (ii) o percentual de acertos no item situado no ponto anterior ao em análise fosse menor que 65%; (iii) o percentual de acertos do item no ponto em análise fosse maior ou igual a 65%; e (iv) o ajuste da TRI fosse bom. Essa metodologia afrouxou os parâmetros para a definição de itens âncoras, permitindo que uma maior quantidade de itens fosse incluída na escala de proficiência (OLIVEIRA, 2008).

Quanto à seleção de pontos para a interpretação, com a justificativa de que em edições anteriores não era possível atribuir significado pedagógico às médias que se situavam entre dois pontos interpretados na escala, foi adotada uma forma de descrição que abrangesse não somente pontos, mas intervalos da escala, denominados níveis de desempenho (INEP, s.d). Esses níveis são apresentados no Quadro 4 a seguir.

Quadro 4 - Definição dos níveis de desempenho da escala de proficiência do ciclo Saeb 1999

Língua portuguesa		**Matemática**	
Nível	**Pontos na escala de proficiência**	**Nível**	**Pontos na escala de proficiência**
150	≥ 150 e < 200	160	≥ 160 e < 175
200	≥ 200 e < 250	175	≥ 175 e < 225
250	≥ 250 e < 300	225	≥ 225 e < 275
300	≥ 300 e < 350	275	≥ 275 e < 325
350	≥ 350 e < 400	325	≥ 325 e < 375
		375	≥ 375 e < 425
		425	≥ 425 e < 475

Fonte: Inep *apud* Horta Neto (2006).

A interpretação desses níveis seguiu a metodologia utilizada em 1995 e 1997, com a submissão dos itens a especialistas, que descreviam o que os indivíduos com proficiências em determinado intervalo sabiam e eram capazes de fazer. No entanto, houve mudança no formato de publicação da descrição da escala. No relatório técnico do Saeb 1999, as descrições de cada nível ocuparam uma página dos anexos, seguidas de um item alocado no respectivo nível, a título de exemplificação. A Figura 6 a seguir mostra uma dessas páginas. Novamente, observa-se o uso de

marcadores textuais para elencar as habilidades, entretanto, os verbos são utilizados no infinitivo e há textos corridos que também descrevem o que os indivíduos demonstraram ser capazes de realizar.

Figura 6 - Trecho da escala de proficiência de língua portuguesa do ciclo Saeb 1999

Língua Portuguesa

Nível 200 ≤ θ < 250

Leitura com compreensão global de textos pequenos, com frases curtas em ordem direta, vocabulário e temática próximos da realidade do aluno.

Os alunos resolvem problemas de leitura a partir da compreensão global do texto, incluindo inferências, localizam informações secundárias, reconstroem uma narrativa, encadeando vários fatos na ordem de aparição, e reconhecem efeitos de sentido de recursos variados (repetição, substituição, onomatopéia).

Na compreensão dos textos, o aluno demonstra:

- Resolver problemas de leitura a partir da compreensão global do texto.
- Localizar informações secundárias.
- Fazer inferências a partir da leitura global.
- Reconstruir uma narrativa, encadeando vários fatos na ordem de aparição.
- Correlacionar, em um texto dado, termos, expressões ou idéias que tenham o mesmo referente.
- Reconhecer efeitos de sentido de recursos variados (repetição, substituição, onomatopéia).
- Progredir na leitura articulando partes do texto.
- Excluir, de uma lista, por verificação, informações plausíveis não contidas no texto.
- Reconhecer e comparar paráfrases, identificando, entre várias, a mais apropriada.
- Reconhecer gêneros textuais a partir da enumeração de características.

Fonte: Extraído do Relatório Saeb 1999, site do Inep (s.d.).

Em 2001, mais uma vez, houve alteração nos níveis da escala de proficiência. Além da inclusão de pontos ou intervalos, os níveis foram nomeados com os números de 1 a 8 em língua portuguesa e 1 a 10 em matemática. O Quadro 5, a seguir, mostra essas informações. Quanto aos

critérios de ancoragem dos itens, segundo relatório desse ciclo, "um item é considerado âncora em um determinado nível quando: (i) o percentual de acerto do item, no nível considerado e nos níveis acima dele, é maior que 65%; (ii) o percentual de acerto do item, nos níveis anteriores, é menor que 65%" (INEP, 2002a, p. 14).

Quadro 5 - Definição dos níveis de desempenho da escala de proficiência do ciclo Saeb 2001

Língua portuguesa		**Matemática**	
Nível	**Pontos na escala de proficiência**	**Nível**	**Pontos na escala de proficiência**
1	≥ 125 e < 150	1	≥ 125 e < 150
2	≥ 150 e < 175	2	≥ 150 e < 175
3	≥ 175 e < 200	3	≥ 175 e < 200
4	≥ 200 e < 250	4	≥ 200 e < 250
5	≥ 250 e < 300	5	≥ 250 e < 300
6	≥ 300 e < 350	6	≥ 300 e < 350
7	≥ 350 e < 375	7	≥ 350 e < 375
8	≥ 375	8	≥ 375 e < 400
		9	≥ 400 e < 425
		10	≥ 425

Fonte: Inep *apud* Horta Neto (2006).

A escala interpretada ganhou um formato diferente em 2001. Em cada nível, as habilidades dos estudantes foram distribuídas de acordo com os tópicos da matriz de referência do Saeb. Especificamente em língua portuguesa, eram seis: procedimentos de leitura; implicações do suporte, do gênero e/ou do enunciador na compreensão dos textos; relação entre textos; coesão e coerência no processamento do texto; relações entre recursos expressivos e efeitos de sentido; e variação linguística (INEP, 2002a). Visualmente também organizada em um quadro, em 2001, foram dispensados os marcadores textuais e surgiram textos descritivos para cada um desses tópicos, que mostram as habilidades que os estudantes demonstraram possuir, indicadas por verbo na 3ª pessoa do plural no presente. Ademais, foram incluídos os percentuais de estudantes de cada série avaliada posicionados em cada nível, à direita. A Figura 7 mostra um trecho dessa escala.

Figura 7 - Trecho da escala de proficiência de língua portuguesa do ciclo Saeb 2001

Nível	Descrição dos Níveis da Escala	Resultados do Saeb 2001
4 200	**Procedimentos de Leitura** Os alunos são capazes de identificar a descrição de um lugar em textos publicitários de revistas e jornais. Identificam tema de texto poético de baixa complexidade como, por exemplo, um poema descritivo. Percebem também o sentido de uma expressão de uso corrente em textos informativos. Identificam informação implícita e o tema em narrativa curta (fábula), especialmente com base em material ilustrativo. Os alunos de 8ª série do E.F., além das habilidades descritas, identificam e interpretam informações contidas em gráficos e tabelas, inferem informações implícitas e identificam o tema em textos poéticos. Os alunos distinguem fato da opinião relativa a esse fato em narrativa histórica. **Coerência e Coesão no Processamento do Texto** Os alunos estabelecem relações anafóricas, isto é, relações entre palavras e suas substituições pronominais, em textos curtos e simples. Reconhecem elementos constitutivos de narrativas: espaço (em crônicas), personagens, conflito e desfecho (em histórias infantis). Os alunos de 8ª série do E.F., além disso, estabelecem relação entre uma palavra de sentido mais genérico e outra de sentido mais específico (relação parte/todo). Estabelecem, também, relações entre partes de um texto, identificando repetições/substituições que contribuem para a sua continuidade. Sabem reconhecer, ainda, relações de causa e conseqüência em textos poéticos. **Relações entre Recursos Expressivos e Efeitos de Sentido** Os alunos identificam o efeito de sentido decorrente da disposição gráfica das palavras em um texto. **Variação Lingüística** Os alunos identificam marcas lingüísticas próprias de textos comerciais. Os alunos de 8ª série do E.F. acrescentam também a habilidade de identificar marcas lingüísticas que evidenciam o locutor e o interlocutor em textos informativos.	*Percentual de alunos entre os Níveis 200 a 250* 4ª S E.F. 18,85 8ª S E.F. 36,37 3ª S E.M. 28,67

Fonte: Inep (2002).

Ainda em 2001, foi feita uma nova tentativa no sentido de associar os pontos da escala aos momentos da aprendizagem formal. No entanto, de acordo com Horta Neto (2006, p. 80), apesar do critério adotado ser mais coerente que o de 1997, visto que utilizava como foco de análise o final dos ciclos de aprendizagem, "novamente, não foi apresentada uma justificativa técnica que permitisse entender os critérios utilizados para estabelecer essas relações". Essa associação é mostrada no Quadro 6 a seguir.

Quadro 6 - Relação entre os níveis da escala e as etapas de ensino proposta no relatório técnico do Saeb 2001

Série	Pontos na escala	
	Língua portuguesa	Matemática
4ª	≥ 125 e < 300	≥ 125 e < 350
8ª	≥ 150 e < 375	≥ 200 e < 400

Fonte: Inep *apud* Horta Neto (2006).

No ciclo de 2003 do Saeb, assim como nos ciclos anteriores, a apresentação dos resultados foi novamente modificada. Apesar de não serem fornecidos argumentos técnicos consistentes que justificassem essa alteração, os níveis da escala de desempenho foram agrupados em quatro classificações – Muito crítico, Crítico, Intermediário e Adequado –, denominadas estágios de construção de competências. Ademais, foram estabelecidas proficiências mínimas por área e etapa da educação básica (INEP, 2006). Essas informações constam dos Quadros 7 e 8 a seguir. Para a ancoragem dos itens nesses pontos, foram adotados os mesmos critérios do ciclo de 2001.

Quadro 7 - Estágios de construção de competência e os pontos da escala de proficiência

Estágios	Pontos na escala			
	Língua portuguesa		Matemática	
	4ª série	8ª série	4ª série	8ª série
Muito crítico	< 125	< 175	< 125	< 200
Crítico	≥ 125 e < 175	≥ 175 e < 250	≥ 125 e < 175	≥ 200 e < 275
Intermediário	≥ 175 e < 250	≥ 250 e < 325	≥ 175 e < 250	≥ 275 e < 375
Adequado	≥ 250 e < 300	≥ 325 e < 375	≥ 250 e < 350	≥ 375 e < 400

Fonte: Inep *apud* Horta Neto (2006).

Quadro 8 - Proficiência mínima por área e etapa da educação básica

Série	Proficiência mínima satisfatória (pontos da escala)	
	Língua portuguesa	Matemática
4ª	200	200
8ª	300	300

Fonte: Inep *apud* Horta Neto (2006).

Quanto à interpretação da escala, no relatório do Saeb 2003 não foi apresentada uma descrição única, por área, como pode ser observado em ciclos anteriores. Dessa vez, para cada série, foram incluídos quadros que detalham as habilidades que os indivíduos demonstraram possuir, organizadas conforme os tópicos das respectivas matrizes de referência. A Figura 8 mostra um trecho da escala de uma das séries avaliadas.

Figura 8 - Trecho da escala de desempenho de língua portuguesa da 4ª série do ensino fundamental do ciclo Saeb 2003

Competência avaliada	Estágio crítico – 36,75%		Estágio intermediário – 39,71%			Estágio adequado – 4,81%		
	Nível 125 17,10%	Nível 150 19,65%	Nível 175 18,96%	Nível 200 13,17%	Nível 225 7,58%	Nível 250 3,27%	Nível 275 1,05%	Nível 300 0,49%
Tópico I								
Localizar informações explícitas em um texto	Localizam informações que completam literalmente o enunciado das questões propostas, em textos curtos, tais como o do gênero "conto" e "convite".							
	Localizam informações explícitas no texto, operando com maior complexidade discursiva, nas seguintes situações: – na identificação da informação em respostas facilitadas pela presença de uma palavra similar (medroso/medo) no texto básico; – na identificação e na comparação de informações para responder ao solicitado; – na identificação e na seleção de informações concorrentes (escolher um nome de personagem entre muitos outros, por exemplo);							
			Localizam informações em textos mais complexos, completando o enunciado, além de operar por meio de paráfrases. [13] Nos trechos em que o discurso direto esteve presente, demonstram êxito na localização de informações literais apresentadas em ordem inversa no texto. Localizam também informações explícitas com base no conjunto de elementos presentes no texto. Demonstram processar, ainda, informações em um nível de abstração mais elevado, em histórias infantis. Localizam informações em texto não-verbal – sem o apoio da escrita –, as temáticas de natureza científica e a prosa poética, com complexidade do tema e do vocabulário. Novos gêneros textuais passam a incorporar o repertório de leitura dos alunos neste nível – matérias de jornal, poemas mais longos e textos enciclopédicos –, além dos contos, destacando-se a presença de variedade lingüística e de temática pouco comum.				
				Localizam informações operando com paráfrases, em história infantil com múltiplas personagens e estrutura discursiva interacional. Identificam, ainda, informações mais detalhadas em contextos conversacionais em narrativas infantis.				
					Localizam informações solicitadas a respeito de um fato, desprezando as informações secundárias que possam competir com a informação mais relevante do texto básico. Localizam informações em texto instrucional, com complexidade vocabular elevada.			

[13] Paráfrase, no contexto desse descritor, é a reprodução das idéias de um trecho do texto, de forma explícita. Portanto, a informação solicitada não se apresenta de forma literal.

Fonte: Inep (2006).

Nesse formato, se percebe ainda o posicionamento das habilidades conforme o ponto da escala e os estágios de construção de competências adotados nesse ciclo, além da indicação dos percentuais de indivíduos em cada um desses subgrupos. Ademais, se nota que as descrições também são iniciadas por verbo na 3ª pessoa do plural no presente, como já utilizado em ciclos anteriores.

Em 2005, foi publicada pelo MEC a portaria n. 931, que trouxe novidades para o Saeb, passando esse a ser constituído por duas avaliações: a Avaliação Nacional da Educação Básica (Aneb) e a Avaliação Nacional do Rendimento Escolar (Anresc), conhecida popularmente como Prova Brasil. A primeira manteve "os objetivos, características e procedimentos da avaliação da educação básica efetuada até agora pelo Saeb realizado por meio de amostras da população" (BRASIL, 2005, Art. 1º, § 1º). Avalia, portanto, os sistemas de ensino público e particular a cada dois anos, com vistas a produzir informações sobre o desempenho dos alunos dos ensinos fundamental e médio.

A segunda, a Prova Brasil, passou a ter como objetivo principal oferecer informações sistemáticas sobre os desempenhos dos estudantes de cada unidade escolar da rede pública de ensino avaliada, tendo como característica o caráter censitário da avaliação. Assim, se pretendia "avaliar a qualidade do ensino ministrado nas escolas, de forma que cada unidade escolar receba o resultado global (BRASIL, 2005, Art. 1º, § 2º).

Esse movimento pode ser entendido como um esforço do Estado em promover a testagem dos estudantes brasileiros e, assim, investir na avaliação como ferramenta de medida da qualidade, da equidade e da eficiência dos sistemas e redes de ensino brasileiros. No entanto, a partir desse ano, o mesmo esforço não foi observado no que se refere à transparência dessa política e à divulgação dos seus resultados. Apenas em 2018 foi publicado um relatório que pretendeu exibir um panorama da década 2005-2015 da Aneb e da Anresc. Essa percepção é corroborada por Horta Neto (2013, p. 153), que em sua tese afirma que

> [...] Desde o ciclo de 2005, não são mais publicados os relatórios técnicos, e as informações relativas às proficiências das escolas passaram a ser divulgadas no site do Inep. Um fato como esse prejudica bastante a transparência de um procedimento de tamanha relevância, que envolve tantas pessoas. Qualquer que seja a pesquisa, uma das regras básicas é o acesso à metodologia utilizada e às informações que deram origem às medidas divulgadas. Portanto, seria importante que o Inep voltasse a produzir e a divulgar os relatórios técnicos.

Apesar dessa constatação, o relatório técnico publicado em 2018, que abrangeu seis edições da Prova Brasil, foi o primeiro a trazer, especificamente na descrição da escala de desempenho, uma metodologia de caráter pedagógico que orientasse a forma como os especialistas deveriam fazer as interpretações dos itens posicionados na escala para além do uso de seus conhecimentos técnicos na área. Segundo o relatório,

> [...] Na metodologia adotada pelo Inep para construção da escala interpretada, os itens são descritos pedagogicamente de acordo com três elementos estruturais: 1) Operação cognitiva: traduz as ações requeridas ao participante do teste para resolver a situação-problema proposta pelo item; 2) Objeto do conhecimento: refere-se aos conhecimentos escolares solicitados ou mobilizados no item para que o respondente execute a operação cognitiva visando à resolução do item; 3) Contexto: considera as situações envolvidas no problema construído pelo item (INEP, 2018, p. 47).

Quanto aos critérios de ancoragem, o relatório não traz informações novas, o que permite supor que são os mesmos adotados nos ciclos de 2001 e 2003. Já no que se refere aos níveis da escala, esses não sofreram qualquer alteração no período 2005-2015. Da mesma maneira que em 2003, a escala de proficiência do período 2005-2015 é separada por etapa de ensino. A Figura 9 a seguir traz um trecho da escala de língua portuguesa para o 5º ano do ensino fundamental.

Figura 9 - Trecho da escala de desempenho de língua portuguesa do 5º ano do ensino fundamental do ciclo Saeb 2005-2015

5° ANO	
Nível	**Descrição das habilidades desenvolvidas**
Nível 0 Desempenho menor que 125	A Prova Brasil não utilizou itens que avaliam as habilidades desse nível. Os estudantes localizados abaixo do nível 125 requerem atenção especial, pois não demonstram sequer habilidades muito elementares.
Nível 1 Desempenho maior ou igual a 125 e menor que 150	Os estudantes provavelmente são capazes de: localizar informações explícitas em textos narrativos curtos, informativos e anúncios. Identificar o tema de um texto. Localizar elementos, como o personagem principal. Estabelecer relação entre partes do texto: personagem e ação; ação e tempo; ação e lugar.
Nível 2 Desempenho maior ou igual a 150 e menor que 175	Além das habilidades anteriormente citadas, os estudantes provavelmente são capazes de: localizar informações explícitas em contos. Identificar o assunto principal e a personagem principal em reportagem e em fábulas. Reconhecer a finalidade de receitas, manuais e regulamentos. Inferir características de personagens em fábulas. Interpretar linguagem verbal e não verbal em tirinhas.
Nível 3 Desempenho maior ou igual a 175 e menor que 200	Além das habilidades anteriormente citadas, os estudantes provavelmente são capazes de: localizar informação explícita em contos e reportagens. Localizar informação explícita em propagandas com ou sem apoio de recursos gráficos. Reconhecer relação de causa e consequência em poemas, contos e tirinhas. Inferir o sentido de palavra, o sentido de expressão ou o assunto em cartas, contos, tirinhas e histórias em quadrinhos, com o apoio de linguagem verbal e não verbal.
Nível 4 Desempenho maior ou igual a 200 e menor que 225	Além das habilidades anteriormente citadas, os estudantes provavelmente são capazes de: identificar informação explícita em sinopses e receitas culinárias. Identificar assunto principal e personagem em contos e letras de música. Identificar formas de representação de medida de tempo em reportagens. Identificar assuntos comuns a duas reportagens. Identificar o efeito de humor em piadas. Reconhecer sentido de expressão, elementos da narrativa e opinião em reportagens, contos e poemas. Reconhecer relação de causa e consequência e relação entre pronomes e seus referentes em fábulas, poemas, contos e tirinhas. Inferir sentido decorrente da utilização de sinais de pontuação e sentido de expressões em poemas, fábulas e contos. Inferir efeito de humor em tirinhas e histórias em quadrinhos.

Fonte: Inep (2018).

Como se pode notar, nessa versão da interpretação da escala também não são utilizados marcadores textuais e cada nível é descrito por um texto corrido que elencam as habilidades que os estudantes provavelmente possuem, cada uma iniciada por verbo no infinitivo. Cabe observar algumas informações inseridas nesse formato. Primeiramente, no nível 0, que representa desempenhos menores que 125, a escala traz as frases “A Prova Brasil não utilizou itens que

avaliam as habilidades desse nível" e "Os estudantes localizados abaixo do nível 125 requerem atenção especial, pois não demonstram sequer habilidades muito elementares", reforçando uma preocupação com o posicionamento de estudantes nesse nível. Destaca-se, porém, que parece haver certa incoerência na combinação dessas frases, pois, ao afirmar que a prova não utilizou itens desse nível, parece ilógico concluir que os estudantes nele localizados não demonstram habilidades elementares, uma vez que a própria escala parece indicar que não foi possível medi-las.

Além disso, o nível 1 é iniciado pela frase "Os estudantes provavelmente são capazes de", que corrobora o caráter probabilístico da escala de desempenho (ANDRADE; TAVARES; VALLE, 2000). Por fim, a inclusão de "Além das habilidades anteriormente citadas, os estudantes provavelmente são capazes de" no início da descrição dos demais níveis reforça a cumulatividade da escala, isto é, que um estudante posicionado em determinado nível, provavelmente, possui as habilidades descritas nesse nível e aquelas descritas nos níveis anteriores (BEATON; ALLEN, 1992).

No ciclo de 2017, os métodos adotados para a construção da escala de proficiência e a sua interpretação não foram diferentes do descrito no relatório que compreende os ciclos de 2005 a 2015. Os critérios de ancoragem e os níveis da escala adotados são os mesmos e as orientações para a análise dos itens pelos especialistas são postas de forma bastante similar (INEP, 2019). A Figura 10 traz mais uma vez a escala do 5º ano para esse ciclo, com apenas uma observação a ser feita, quanto ao nível 0. Dessa vez, o Inep optou por inserir uma nota que informa que "O Saeb não especifica as habilidades desenvolvidas no nível 0 da escala". No entanto, nenhuma justificativa foi dada quanto ao porquê dessa não especificação.

Figura 10 - Trecho da escala de desempenho de língua portuguesa do 5º ano do ensino fundamental do ciclo Saeb 2017

5° ano	
Nível	**Descrição das habilidades desenvolvidas**
Nível 1* Desempenho maior ou igual a 125 e menor que 150	Os estudantes provavelmente são capazes de: localizar informações explícitas em textos narrativos curtos, informativos e anúncios. Identificar o tema de um texto. Localizar elementos, como o personagem principal. Estabelecer relação entre partes do texto: personagem e ação; ação e tempo; ação e lugar.
Nível 2 Desempenho maior ou igual a 150 e menor que 175	Além das habilidades anteriormente citadas, os estudantes provavelmente são capazes de: localizar informações explícitas em contos. Identificar o assunto principal e a personagem principal em reportagem e em fábulas. Reconhecer a finalidade de receitas, manuais e regulamentos. Inferir características de personagens em fábulas. Interpretar linguagem verbal e não verbal em tirinhas.
Nível 3 Desempenho maior ou igual a 175 e menor que 200	Além das habilidades anteriormente citadas, os estudantes provavelmente são capazes de: localizar informação explícita em contos e reportagens. Localizar informação explícita em propagandas com ou sem apoio de recursos gráficos. Reconhecer relação de causa e consequência em poemas, contos e tirinhas. Inferir o sentido de palavra, o sentido de expressão ou o assunto em cartas, contos, tirinhas e histórias em quadrinhos, com o apoio de linguagem verbal e não verbal.
Nível 4 Desempenho maior ou igual a 200 e menor que 225	Além das habilidades anteriormente citadas, os estudantes provavelmente são capazes de: identificar informação explícita em sinopses e receitas culinárias. Identificar assunto principal e personagem em contos e letras de música. Identificar formas de representação de medida de tempo em reportagens. Identificar assuntos comuns a duas reportagens. Identificar o efeito de humor em piadas. Reconhecer sentido de expressão, elementos da narrativa e opinião em reportagens, contos e poemas. Reconhecer relação de causa e consequência e relação entre pronomes e seus referentes em fábulas, poemas, contos e tirinhas. Inferir sentido decorrente da utilização de sinais de pontuação e sentido de expressões em poemas, fábulas e contos. Inferir efeito de humor em tirinhas e histórias em quadrinhos.

Fonte: Inep (2019)

Como fonte de acesso às escalas de proficiência do Saeb, além dos relatórios técnicos relacionados a cada ciclo, no site do Inep[7] estão disponíveis as escalas mais atualizadas da avaliação. Retiradas do contexto de um relatório técnico sobre o processo avaliativo, as escalas

[7]A título de divulgação para a sociedade, o Inep publica as escalas de proficiência também em seu portal, no seguinte endereço: http://portal.inep.gov.br/web/guest/educacao-basica/saeb/matrizes-e-escalas.

são disponibilizadas à comunidade para fins de consulta. Conforme os ciclos anteriores, elas são organizadas por ano escolar. A Figura 11 traz um trecho da escala do 5º ano de língua portuguesa.

Figura 11 - Trecho da escala de desempenho de língua portuguesa do 5º ano do ensino fundamental retirado do site do Inep

Nível	Descrição do Nível
Nível 0 Desempenho menor que 125	A Prova Brasil não utilizou itens que avaliam as habilidades deste nível. Os estudantes localizados abaixo do nível 125 requerem atenção especial, pois não demonstram habilidades muito elementares.
Nível 1 Desempenho maior ou igual a 125 e menor que 150	Os estudantes provavelmente são capazes de: • Localizar informações explícitas em textos narrativos curtos, informativos e anúncios. • Identificar o tema de um texto. • Localizar elementos como o personagem principal. • Estabelecer relação entre partes do texto: personagem e ação; ação e tempo; ação e lugar.
Nível 2 Desempenho maior ou igual a 150 e menor que 175	Além das habilidades anteriormente citadas, os estudantes provavelmente são capazes de: • Localizar informações explícitas em contos. • Identificar o assunto principal e a personagem principal em reportagem e em fábulas. • Reconhecer a finalidade de receitas, manuais e regulamentos. • Inferir características de personagens em fábulas. • Interpretar linguagem verbal e não-verbal em tirinhas.
Nível 3 Desempenho maior ou igual a 175 e menor que 200	Além das habilidades anteriormente citadas, os estudantes provavelmente são capazes de: • Localizar informação explícita em contos e reportagens. • Localizar informação explícita em propagandas com ou sem apoio de recursos gráficos. • Reconhecer relação de causa e consequência em poemas, contos e tirinhas. • Inferir o sentido de palavra, o sentido de expressão ou o assunto em cartas, contos, tirinhas e histórias em quadrinhos com o apoio de linguagem verbal e não verbal.
Nível 4 Desempenho maior ou igual a 200 e menor que 225	Além das habilidades anteriormente citadas, os estudantes provavelmente são capazes de: • Identificar informação explícita em sinopses e receitas culinárias. • Identificar assunto principal e personagem em contos e letras de música. • Identificar formas de representação de medida de tempo em reportagens. • Identificar assuntos comuns a duas reportagens. • Identificar o efeito de humor em piadas. • Reconhecer sentido de expressão, elementos da narrativa e opinião em reportagens, contos e poemas. • Reconhecer relação de causa e consequência e relação entre pronomes e seus referentes em fábulas, poemas, contos e tirinhas. • Inferir sentido decorrente da utilização de sinais de pontuação e sentido de expressões em poemas, fábulas e contos. • Inferir efeito de humor em tirinhas e histórias em quadrinhos.

Fonte: Extraído do site do Inep (2020).

Novamente, se percebe que os níveis de desempenho não foram alterados, mas chama a atenção a forma em que a escala interpretada foi organizada, retomando o uso de marcadores textuais para indicar as habilidades posicionadas em cada nível. Em todos os outros aspectos (uso de verbos no infinitivo e as frases no início de cada nível), a escala mantém a estrutura já observada no ciclo anterior.

Considerações finais

O presente estudo elucida os pressupostos metodológicos que embasam a definição e construção de escalas de proficiência, a partir do levantamento de pesquisas em avaliação educacional e teorias de medida. Utilizando no referencial teórico autores ligados à pesquisa em educação, psicologia e estatística, foi possível esclarecer os procedimentos e parâmetros envolvidos na construção de escalas. Essa investigação foi fundamental para embasar a análise das escalas de proficiência do Saeb, objetivo maior deste estudo.

Visando atender a esse objetivo, foi feita uma análise documental dos relatórios técnicos divulgados pelo Inep concernentes aos ciclos de avaliação do Saeb no período de 1995 a 2017. A leitura desse material revelou que, entre os anos de 1995 e 1999, em que foram realizados três ciclos da avaliação, a comparação de resultados foi inviabilizada pelo fato de terem sido realizadas modificações nos pontos escolhidos para a interpretação da escala (HORTA NETO, 2006). Essas alterações, apesar de terem o objetivo de aprimorar a metodologia utilizada, garantindo maior confiabilidade ao processo, dificultam "o trabalho de um gestor que procure entender o que vem acontecendo com o sistema educacional pelo qual é responsável" (HORTA NETO, 2006, p. 74). Um dos grandes méritos da avaliação em larga escala é permitir verificar, durante determinado período, pontos de evolução ou involução e, a partir daí, criar estratégias para reforçar aspectos positivos e corrigir os negativos. No entanto, entre 1995 e 1999 foi inviável fazer esse diagnóstico, tendo em vista que em cada ciclo foram escolhidos pontos diferentes para a ancoragem dos itens e para a interpretação da escala.

Por outro lado, desde 1999, quando foram estabelecidos os níveis de desempenho nos intervalos entre dois pontos da escala, a metodologia adotada pelo Inep para a construção de

escalas de proficiência parece ter se consolidado. Entre 1999 e 2001 ainda são observados alguns ajustes, mas em 2003 é instituída a métrica com intervalos de 25 pontos (meio desvio padrão), utilizada até hoje (INEP, 2006). Essa permanência é fundamental para a definição de uma série histórica do desempenho em uma avaliação que, por sua vez, fornecerá indícios sobre o desenvolvimento do processo de ensino-aprendizagem no país.

Outro ponto que chama a atenção é o fato de ter sido publicado apenas um relatório no período entre 2005 e 2015, o que evidencia que, mesmo com todos os investimentos e esforços no sentido de garantir a testagem dos estudantes brasileiros, pouco tem sido feito quanto à publicação dos resultados dessas avaliações (HORTA NETO, 2013). Entendendo a escala de proficiência como uma ferramenta que visa dar subsídios a professores e gestores a respeito do desempenho de seus estudantes, é primordial que eles tenham acesso ao seu conteúdo, compreendam o seu significado e, principalmente, tenham as informações necessárias para melhorar o seu trabalho em sala de aula e, consequentemente, a qualidade do ensino no Brasil. Se essa devolutiva não acontece, perde-se o sentido em realizar a própria avaliação. Como afirma Horta Neto (2013, p. 94), o uso eficaz desses resultados deve partir do pressuposto de que

> [...] compreender melhor as diferenças de complexidade existentes entre as diferentes habilidades testadas [...] pode trazer informações importantes sobre as limitações existentes, que impedem o avanço na aprendizagem. Determinar que a proficiência está associada a um número ou que ela se encontra em determinado nível da escala de proficiência de nada adianta. É preciso dar significado pedagógico ao número, pois, afinal, qual o sentido de se afirmar que a proficiência média dos alunos de uma escola encontra-se no nível 2?

Finalmente, sobre os resultados do Saeb, é importante frisar a necessidade de se manter uma constância no seu formato de divulgação entre os ciclos e garantir produtos que de fato estejam ao alcance da comunidade. A avaliação em larga escala e todas as metodologias que a circundam são bastante complexas e "dispõem de tecnologia que parece não ter sido suficientemente sistematizada e difundida no cenário educacional, e em particular, no cenário brasileiro" (FONTANIVE; ELLIOT; KLEIN, 2007). Esclarecer seus pressupostos em uma linguagem acessível e concisa é vital para garantir que o seu objetivo seja alcançado. É preciso ampliar o rol de produtos de divulgação de uma política tão importante como essa, mas também garantir que a comunidade acadêmica tenha informações necessárias para desenvolver estudos que

possam não só melhorar a avaliação, mas também proporcionar maior conhecimento dos seus resultados e consequências.

Referências

ANDRADE, D. F. de; TAVARES, H. R.; VALLE, R. da C. Teoria da Resposta ao Item: conceitos e aplicações. São Paulo: Associação Brasileira de Estatística, 2000. Disponível em: <https://docs.ufpr.br/~aanjos/CE095/LivroTRI_DALTON.pdf>. Acesso em: 21 ago. 2021.

ARAUJO, E. A. C. de; ANDRADE, D. F. de; BORTOLOTTI, S. L. V. Teoria da Resposta ao Item. **Rev Esc Enferm USP**, São Paulo, v. 43, n. esp., p. 1000-1008, 2009. Disponível em: <https://www.scielo.br/j/reeusp/a/VS9FdSVm6CsSxQYkJ5nr8tD/?lang=pt>. Acesso em: 21 ago. 2021.

BEATON, A. E.; ALLEN, N. L. Interpreting scales through scale anchoring. **Journal of Educational Statistics**, v. 17, p. 191-204, 1992. Disponível em: <https://www.jstor.org/stable/1165169>. Acesso em: 21 ago. 2021

BRASIL. Lei n. 9.394/1996, de 20 de dezembro de 1996. Estabelece as diretrizes e bases da educação nacional. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l9394.htm>. Acesso em: 13 mai. 2020.

BRASIL. Portaria n. 931, de 21 de março de 2005. Institui o Sistema de Avaliação da Educação Básica - SAEB. **Diário Oficial da União**, Brasília, n. 55, p. 17, 22 mar. 2005. Seção 1. Disponível em: <https://download.inep.gov.br/educacao_basica/prova_brasil_saeb/legislacao/Portaria931_NovoSaeb.pdf>. Acesso em: 21 ago. 2021.

CALDERÓN, A. I.; BORGES, R. M. Avaliação em larga escala na educação básica: usos e tensões teórico-epistemológicas. **Meta: Avaliação**, Rio de Janeiro, v. 12, n. 34, p. 28-58, jan./mar. 2020. Disponível em: <http://dx.doi.org/10.22347/2175-2753v12i34.2281>. Acesso em: 21 ago. 2021.

FONTANIVE, N. S.; ELLIOT, L. G.; KLEIN, R. Os desafios da apresentação dos resultados da avaliação de sistemas escolares a diferentes públicos. **REICE - Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación**, Madri, v. 5, n. 2, p. 262-273, 2007. Disponível em: <http://hdl.handle.net/10486/660969>. Acesso em: 21 ago. 2021.

HAMBLETON, R. K.; SWAMINATHAN, H.; ROGERS, H. J. **Fundamentals of Item Response Theory**. Newbury Park: Sage University Paper, 1991. Disponível em: <https://doi.org/10.1177/014662169301700309>. Acesso em: 21 ago. 2021.

HORTA NETO, J. L. **Avaliação externa**: a utilização dos resultados do SAEB 2003 na gestão do sistema público de Ensino Fundamental no Distrito Federal. Brasília, 2006. 144 f. Dissertação (Mestrado) – Faculdade de Educação, Universidade de Brasília, 2006. Disponível em: <https://repositorio.unb.br/handle/10482/5811>. Acesso em: 21 ago. 2021.

HORTA NETO, J. L. **As avaliações externas e seus efeitos sobre as políticas educacionais**: uma análise comparada entre a União e os Estados de Minas Gerais e São Paulo. Brasília, 2013. 358 f. Tese (Doutorado em Política Social) - Programa de Pós-Graduação em Política Social da Universidade de Brasília, 2013. Disponível em: <https://repositorio.unb.br/handle/10482/14398>. Acesso em: 21 ago. 2021.

INEP. **Resultados do Saeb/95**: escalas de proficiência. 2. ed. Brasília: Inep, 1998. Disponível em: <http://inep.gov.br/documents/186968/484421/Resultados+do+SAEB-95+escalas+de+profici%C3%AAncia/5367fe05-c42e-4aeb-a771-ded2af158322?version=1.2>. Acesso em: 21 ago. 2021.

INEP. **Saeb 97**: primeiros resultados. Brasília: Inep, 1999. Disponível em: <https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/saeb_97_primeiros_resultados.pdf>. Acesso em: 21 ago. 2021.

INEP. **Saeb**: resultados 99. Brasília: Inep, s.d. Disponível em: <http://inep.gov.br/informacao-da-publicacao/-/asset_publisher/6JYIsGMAMkW1/document/id/485776>. Acesso em: 21 ago. 2021.

INEP. **Relatório Saeb 2001**: língua portuguesa. Brasília: Inep, 2002a. Disponível em: <http://www.dominiopublico.gov.br/download/texto/me0000131.pdf>. Acesso em 21 ago. 2021.

INEP. **Relatório Saeb 2001**: matemática. Brasília: Inep, 2002b. Disponível em: <https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/relatorio_saeb_2001_matematica.pdf>. Acesso em: 21 ago. 2021.

INEP. **Relatório Nacional Saeb 2003**. Brasília: Inep, 2006. Disponível em: <http://portal.inep.gov.br/informacao-da-publicacao/-/asset_publisher/6JYIsGMAMkW1/document/id/489262>. Acesso em: 21 ago. 2021.

INEP. **Relatório Saeb (Aneb e Anresc) 2005-2015**: panorama da década. Brasília: Inep, 2018. Disponível em: <http://portal.inep.gov.br/informacao-da-publicacao/-/asset_publisher/6JYIsGMAMkW1/document/id/1473828>. Acesso em: 21 ago. 2021.

INEP. **Relatório Saeb**. Brasília: Inep, 2019. Disponível em: <http://portal.inep.gov.br/informacao-da-publicacao/-/asset_publisher/6JYIsGMAMkW1/document/id/6730262>. Acesso em: 21 ago. 2021.

KISTEMANN JR, M. A.; GOUVÊA, C. de L. Uma investigação com professores de matemática e sua leitura dos resultados das avaliações em larga escala (Proeb). **Revista Pesquisa e Debate**

em Educação, Juiz de Fora, v. 9, n. 1, p. 606-624, 2019. Disponível em: <https://doi.org/10.34019/2237-9444.2019.v9.31132>. Acesso em: 21 ago. 2021.

KLEIN, R.; FONTANIVE, N. S. Avaliação em larga escala: uma proposta inovadora. **Em Aberto**, Brasília, ano 15, n. 66, p. 29-34, abr./jun. 1995. Disponível em: <https://doi.org/10.24109/2176-6673.emaberto.15i66.%25p>. Acesso em: 21 ago. 2021.

OLIVEIRA, L. K. M de. **Três investigações sobre escalas de proficiência e suas interpretações**. Rio de Janeiro, 2008. 216 f. Tese (Doutorado em Educação) – Departamento de Educação, Pontifícia Universidade Católica do Rio de Janeiro, 2008. Disponível em: <http://www.dominiopublico.gov.br/pesquisa/DetalheObraForm.do?select_action=&co_obra=125579>. Acesso em: 21 ago. 2021.

PASQUALI, L.; PRIMI, R. Fundamentos da Teoria da Resposta ao Item. **Avaliação Psicológica**, Campinas, v. 2, n. 2, p. 99-110, 2003. Disponível em: <http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1677-04712003000200002>. Acesso em: 21 ago. 2021.

PERRY, F. A. **Escalas de proficiência**: diferentes abordagens de interpretação na avaliação educacional em larga escala. Juiz de Fora, 2009. 119 f. Dissertação (Mestrado em Educação) - Programa de Pós-Graduação em Educação, Universidade Federal de Juiz de Fora, 2009. Disponível em: <https://repositorio.ufjf.br/jspui/handle/ufjf/3835>. Acesso em: 21 ago. 2021.

SALKIND, N. J. (ed.) **Encyclopedia of measurement and statistics**. v. 1. Thousand Oaks: SAGE Publications, 2007. Disponível em: <https://dx.doi.org/10.4135/9781412952644>. Acesso em: 21 ago. 2021.

SILVA, F. S. da; LEAL, T. F. Escala de proficiência da Prova Brasil: o que informa aos professores? **Revista Leia Escola**, Campina Grande, v. 18, n. 3, p. 90-108, 2018. Disponível em: <http://dx.doi.org/10.35572/rle.v18i3.1057>. Acesso em: 21 ago. 2021.

SOUSA, S. Z. L.; OLIVEIRA, R. P. de. Sistemas estaduais de avaliação: uso dos resultados, implicações e tendências. **Cadernos de Pesquisa**, São Paulo, v. 40, n. 141, p. 793-822, set./dez. 2010. Disponível em: <https://www.scielo.br/j/cp/a/HfYnBHFv4x63bWY6nkfJt7H/abstract/?lang=pt>. Acesso em: 21 ago. 2021.

VALLE, R. da C. Teoria de Resposta ao Item. **Estudos em Avaliação Educacional**, São Paulo, n. 21, p. 7-92, 2000. Disponível em: <https://doi.org/10.18222/eae02120002225>. Acesso em: 21 ago. 2021.

ESTUDO II

O ERRO EM AVALIAÇÕES EM LARGA ESCALA: COMO ANALISÁ-LOS E INTERPRETÁ-LOS?

Introdução

A tomada de decisões, em qualquer esfera, pressupõe uma avaliação. É preciso entender a realidade antes de decidir sobre quais rumos deverão ser seguidos ou reorientados. Avaliar, por sua vez, implica na realização de um diagnóstico. Segundo Luckesi (2000, s.p.), diagnosticar

> [...] constitui-se de uma constatação e de uma qualificação do objeto da avaliação. [...] A constatação sustenta a configuração do 'objeto', tendo por base suas propriedades, como estão no momento. O ato de avaliar, como todo e qualquer ato de conhecer, inicia-se pela constatação, que nos dá a garantia de que o objeto é como é. Não há possibilidade de avaliação sem a constatação.

Com base nessa constatação, qualifica-se, atribui-se um significado ao objeto de avaliação. (LUCKESI, 2000). Esses conceitos aplicam-se também às avaliações de desempenho dos estudantes feitas ao longo da sua trajetória escolar. É com o objetivo de se conhecer a realidade do estudante em sala de aula ou de um sistema de ensino, por exemplo, que se realizam avaliações, sejam as de acompanhamento do desenvolvimento curricular, também chamadas internas, no primeiro caso, ou as externas nacionais, no segundo.

Compreende-se, a partir dessa afirmação, que, independentemente do alcance e do objetivo de uma avaliação, o foco deverá recair sobre o diagnóstico, que determinará a decisão a ser tomada para melhorar a qualidade do ensino. Na construção desse diagnóstico, geralmente realizam-se testes, que têm como objetivo medir as competências ou habilidades de um estudante. Em avaliações externas, que são o objeto de estudo da presente pesquisa, nas técnicas usualmente empregadas, o foco de análise recai sobre os acertos dos estudantes nesses testes. É com base neles

que se calcula a nota que irá representar, na prática, o quanto se acredita que aquele estudante sabe a respeito de algum assunto ou conteúdo.

Dessa forma, apesar de os erros terem grande impacto, raramente são interpretados de modo que se entenda o que significam pedagogicamente em relação ao desempenho. Os erros são considerados (ou desconsiderados) para composição das notas, mas será que informam ao estudante em quais aspectos ele deve melhorar ou em quais deve concentrar seus estudos para superá-los?

A valorização do erro como estratégia didática para compreender o desempenho dos estudantes e contribuir para a prática pedagógica não é temática nova (SCHUBRING, 1998). No entanto, ainda é possível identificar na literatura recente sobre o tema uma preocupação em conscientizar professores sobre a importância desse recurso (GONÇALVES, 2014; SPINILLO et al, 2014). Sobre isso, os autores Nogaro e Granella (2004) afirmam que:

> [...] é necessidade urgente que escola e educadores entendam que o erro, na aprendizagem, é a manifestação de uma conduta não aprendida, que emerge a partir de um padrão de conduta cognitivo, e que serve de ponto de partida para o avanço, na medida em que são identificados e compreendidos positivamente, em direção à aprendizagem do aluno, possibilitando a sua correção de forma hábil e inteligente. Isso significa uma forma consciente e elaborada, na conduta docente, em interpretar o erro na aprendizagem como uma possibilidade de crescimento e de valorização do aluno bem como um passo à frente na relação professor-aluno (NOGARO; GRANELLA, 2004, p. 39-40).

A justificativa para a relevância do erro no processo de ensino-aprendizagem está no fato de que, conforme Rosso e Berti (2010, p. 1030-1031) afirmam, “os erros cometidos por uma criança são reveladores do conhecimento construído e das operações executadas pelos sujeitos; o erro não é a negação do conhecer, mas expressão da sua dinâmica própria”. Também de acordo com Gonçalves (2014, p. 29), “não se trata de algo ruim [o erro], mas, de um fenômeno que, naturalmente, faz parte de todo processo de construção do conhecimento, seja ele científico ou escolar”.

A Base Nacional Comum Curricular (BNCC) abarca essa perspectiva ao enfatizar a importância do acolhimento da criança, do adolescente, do jovem e do adulto para promover o

desenvolvimento pleno considerando as suas singularidades e diversidades. Esse documento defende que a

> [...] educação básica deve visar à formação e ao desenvolvimento humano global, o que implica compreender a complexidade e a não linearidade desse desenvolvimento, rompendo com visões reducionistas que privilegiam ou a dimensão intelectual (cognitiva) ou a dimensão afetiva (BRASIL, 2017).

Considerando esse cenário, parte-se do pressuposto de que, conforme afirmam Spinillo et al (2014, p. 7), "interpretar os erros torna possível compreender que estes podem decorrer de formas de raciocinar distintas, umas mais e outras menos elementares". Assim, se torna necessário entender quais são esses distintos raciocínios, pois:

> Ter um olhar diferenciado sobre os erros que os alunos cometem, significa ter um olhar sobre o modo como eles aprendem. A busca pela interpretação do erro do estudante no processo de ensino e aprendizagem é uma maneira de lhe possibilitar a oportunidade de reelaborar seus conhecimentos sobre o objeto de estudo. No entanto, para que isso seja possível é importante que o professor tenha a concepção de que o erro faz parte do processo de construção do conhecimento e não que o erro seja uma simples manifestação da falta de conhecimento ou mesmo de problemas de aprendizagem (GONÇALVES, 2014, p. 28).

Buscando ampliar o alcance dessa proposta de interpretação dos erros cometidos pelos estudantes no processo de desenvolvimento do conhecimento, o presente estudo buscou investigar como essa interpretação poderia ser aplicada ao universo da avaliação em larga escala, mais especificamente, na edição de 2018 da avaliação denominada Prova São Paulo[8], aplicada pelo município de São Paulo aos estudantes da rede pública.

Em 2005, o município de São Paulo instituiu o Sistema de Avaliação de Aproveitamento Escolar dos Alunos da Rede Municipal de Ensino de São Paulo (Portaria nº 2.639, de 10 de março de 2017) com o objetivo de:

> I – reorientação da proposta pedagógica do Ensino Fundamental regular, de modo a aprimorá-la; II – viabilização da articulação dos resultados da avaliação com o planejamento escolar, a formação dos professores e o estabelecimento de metas para o projeto pedagógico de cada escola; III – orientação para os trabalhos desenvolvidos com os estudantes que necessitam de reforço na aprendizagem. (SÃO PAULO, 2017).

[8] O acesso aos dados da Prova São Paulo 2018 foi viabilizado mediante solicitação enviada por e-mail ao Núcleo Técnico em Avaliação da Secretaria Municipal de Ensino de São Paulo, aprovada em 20 de julho de 2020.

Como parte integrante desse sistema, a Prova São Paulo foi aplicada pela primeira vez em 2007, visando avaliar as competências em língua portuguesa (com foco em leitura) e em matemática (com foco na resolução de problemas). Posteriormente, foi inserido o componente ciências naturais. Aplicada anualmente, seus resultados eram expressos na mesma escala que o Saeb, tendo os seguintes níveis de proficiência: Abaixo do básico, Básico, Adequado e Avançado (SOUSA; FERRAROTO, 2016; CHAPPAZ; ALAVARSE, 2017). Tais níveis representam, para língua portuguesa, no 5º ano do ensino fundamental, as proficiências delimitadas no Quadro 9 a seguir.

Quadro 9 - Definição dos níveis de desempenho da escala de proficiência de língua portuguesa do 5º ano do ensino fundamental na Prova São Paulo 2018

Níveis da escala	Pontos na escala de proficiência
Abaixo do básico	135
Básico	≥ a 135 e < 185
Adequado	≥ a 185 e < 235
Avançado	≥ 235

Fonte: Elaborado pela autora (2022).

Utilizando as informações provenientes dessa avaliação, se buscou verificar em que medida é possível associar os erros cometidos por estudantes em um teste à proficiência obtida a partir dos resultados alcançados nesse mesmo teste. Como ferramenta estatística para essa análise, foi utilizada a Teoria de Resposta ao Item (TRI), nos seus modelos politômicos de respostas nominais (MRN) e de respostas graduais (MRG) de Samejima (1969). Em ambos, se considera cada alternativa como uma categoria de resposta. Para o primeiro, pressupõe-se que não existe ordenação entre essas categorias e, no segundo, que essa ordenação existe. A escolha desses modelos foi feita com base no entendimento de que "o conhecimento parcial dos sujeitos se revelaria com maior ou menor intensidade dentre as alternativas incorretas de um item, o que, sempre que ponderado, acarretaria uma maior precisão da medida como um todo" (PINHEIRO; COSTA; CRUZ, 2010, p. 438). O objetivo, portanto, é "maximizar a precisão da habilidade estimada usando toda a informação contida nas respostas dos indivíduos, e não apenas se o item

foi respondido corretamente ou não" (ANDRADE; TAVARES; VALLE, 2000). Isso porque, de acordo com R. Darrell Bock (1972, p.29):

> É sabido que sujeitos que respondem incorretamente a um item de múltipla escolha dificilmente se distribuem uniformemente entre as opções incorretas. Isso sugere que as respostas erradas contêm informações que podem ser aplicadas à estimação da habilidade latente (tradução minha)[9].

O autor defende, portanto, a adoção de modelos que expressem a probabilidade de marcação de cada alternativa do item em função da habilidade do sujeito. Como forma de garantir a extração de informações mais confiáveis, Bock (1972) sugere também a construção de opções erradas plausíveis, considerando o contexto abordado pelo item, o que é tido como fundamental para a elaboração de bons itens de avaliação (HALADYNA; DOWNING; RODRÍGUEZ, 2002; MORENO; MARTÍNEZ; MUÑIZ, 2004; BÉLANGER, 2009; RABELO, 2013; ZIMMARO, 2016). Ademais, de acordo com Pinheiro, Costa e Cruz (2010, p. 439):

> Quando bem formulados, os distratores presentes nos itens nominais costumam indicar diferentes linhas de raciocínio, o que indica vieses de pensamento, vícios de linguagem, limiares cognitivos e, mesmo, algum conhecimento específico de uma parcela da população.

Com base nesse referencial, **o objetivo geral deste estudo é construir uma escala interpretada que leve em consideração o erro do estudante na aplicação de um teste de língua portuguesa**, a partir da qual se possa estabelecer um perfil dos erros cometidos por estudantes com baixa proficiência. Como objetivos específicos, tem-se: (i) modelar cada uma das opções dos itens selecionados para a pesquisa de acordo com os modelos estatísticos adotados; (ii) posicionar as opções erradas em uma mesma métrica, construindo uma escala; e (iii) fazer a interpretação pedagógica dessas opções, gerando evidências de intervenção pedagógica.

Seguindo a trilha metodológica das escalas de proficiência do Saeb, para a definição da escala a que esse estudo se propõe, são utilizados os mesmos procedimentos adotados na construção das escalas do Saeb, porém, dessa vez focalizando os erros e não os acertos dos estudantes. Tais procedimentos envolvem: (i) o estabelecimento de critérios para a ancoragem dos

[9] […] it is well known that subjects who answer a multiple-choice item incorrectly are unlikely to distribute their responses uniformly over the incorrect alternatives. This suggests that "wrong" responses contain information which might be applied to the estimation of latent ability (BOCK, 1972, p. 29).

itens (no caso, das opções erradas); e (ii) seleção dos pontos para a interpretação da escala. Cabe destacar que tais procedimentos foram detalhados no Estudo I da presente dissertação, que explorou os métodos para definição e descrição das escalas de desempenho do Saeb no período de 1995 a 2017. Ademais, optou-se pela utilização desses procedimentos em virtude da escala de proficiência do Saeb ser referência para as avaliações estaduais e municipais conduzidas pelo país (SOUSA; OLIVEIRA, 2010), inclusive a Prova São Paulo.

Com esse estudo, espera-se avançar na pesquisa sobre a interpretação dos erros dos estudantes em testes cognitivos e sobre a utilização de ferramentas estatísticas para essa análise, buscando responder à seguinte pergunta orientadora: *como construir uma escala que leve em consideração o erro do estudante em avaliações em larga escala*?

Método

Para desenvolver esse estudo, foi composta uma base de dados com 91 itens de língua portuguesa aplicados para o 5.º ano do ensino fundamental na edição de 2018 da Prova São Paulo. Como mencionado, nas análises foram utilizados o MRN e o MRG de Samejima, usando como referência a metodologia adotada por Gabriela Thamara de Freitas Barros (2016) em sua dissertação de mestrado, em que buscou testar um conjunto de modelos da TRI para a construção de um indicador de nível socioeconômico. Tal opção se deu em virtude da apresentação detalhada que a autora trouxe da aplicabilidade de diferentes métodos estatísticos na análise de dados educacionais, facilitando assim a pesquisa proposta neste estudo.

Vale esclarecer que esses dois modelos utilizam os parâmetros de discriminação (*a*) e o de dificuldade (*b*), também presentes no modelo logístico de três parâmetros[10] da TRI, o mais comumente utilizado nas análises dos resultados de avaliações em larga escala.

Ressalta-se que o parâmetro *a* indica o quanto o item é capaz de diferenciar os indivíduos em relação ao conhecimento que demonstram possuir quando respondem o item (HAMBLETON; SWAMINATHAN; ROGERS, 1991). Um valor maior que zero indica uma discriminação

[10] Em análises utilizando o modelo logístico de três parâmetros da TRI, cada item apresenta um único valor de a, que representa a discriminação do item, assim como um único valor de b, de dificuldade, e um único valor de c, que representa a probabilidade de acerto casual do item (ANDRADE; TAVARES; VALLE, 2000).

positiva, ou seja, que é provável que um indivíduo que acerta determinado item de fato possua a habilidade avaliada. Quanto mais alto for esse valor, maior é o poder de discriminação do item. Já o parâmetro *b* diz respeito à dificuldade do item, normalmente expressa em uma escala de -3 a 3. Quanto maior for o valor de *b*, maior é a habilidade requerida para alcançar essa probabilidade de acerto. Em contrapartida, quanto menor for o valor de *b*, menor é a habilidade. (HAMBLETON; SWAMINATHAN; ROGERS, 1991).

O MRN foi utilizado nas análises para confirmar o ordenamento das alternativas em relação à habilidade e o gabarito do item em relação ao modelo. Nessa análise, se buscou verificar se, em cada item analisado, havia uma ordenação na probabilidade de resposta em relação à habilidade do indivíduo. O objetivo era avaliar se, dentre aqueles que erravam os itens, havia uma diferenciação no sentido de determinada opção errada ser escolhida por indivíduos com habilidade muito baixa e outra opção também errada ser escolhida por indivíduos com habilidade ainda baixa, porém, maior que a do anterior.

Para tanto, foram calculados os valores do parâmetro *a* de cada opção. Tendo em vista que os itens do *corpus* possuem 4 opções, para atender ao objetivo mencionado no parágrafo anterior, esperava-se que $a_1 < a_2 < a_3 < a_4$, sendo que, quanto mais próximo de 0, menor seria a habilidade do indivíduo e, quanto mais próxima de 3, maior seria essa habilidade. Nesse cenário e considerando um item que pedagogicamente não tenha problemas, o gabarito teria o maior valor de *a*, enquanto os distratores se ordenariam do maior para o menor conforme o nível de habilidade do respondente. Após três rodadas dessa análise, chegou-se aos resultados apresentados na Tabela 1 a seguir. Destaca-se, ainda, que essa ordenação não está relacionada à ordem em que as alternativas do item são apresentadas no teste.

Cabe destacar que nessa etapa de calibração, três questões (36, 47 e 88) não apresentaram o comportamento esperado quanto à ordenação dos valores de *a*. Dada essa não convergência, que pode ser explicada por problemas na elaboração dos itens, optou-se pela exclusão dessas questões das análises, ficando o *corpus* com 88 questões.

Na segunda etapa de análise, os itens foram calibrados segundo o MRG. Diferentemente do MRN, no MRG calcula-se a probabilidade de um indivíduo com determinada habilidade responder determinada alternativa a partir dos parâmetros *a* dos itens e dos *b* das opções. Utilizando os resultados obtidos na calibração com o MRN, deu-se continuidade à análise

mantendo a exclusão das questões 36, 47 e 88. A Tabela 2, a seguir, traz os resultados da modelagem pelo MRG.

Tabela 1 - Parâmetros de calibração segundo o MRN (continua)

Questões	a_1	a_2	a_3	a_4
Q1	0	0,757	0,933	3
Q2	0	0,112	0,863	3
Q3	0	0,959	1,576	3
Q4	0	0,518	1,362	3
Q5	0	0,03	1,636	3
Q6	0	1,885	2,924	3
Q7	0	0,706	0,839	3
Q8	0	1,214	1,474	3
Q9	0	0,649	0,659	3
Q10	0	0,022	0,362	3
Q11	0	0,409	0,791	3
Q12	0	0,023	0,097	3
Q13	0	0,281	1,273	3
Q14	0	0,558	1,209	3
Q15	0	0,816	0,829	3
Q16	0	0,957	1,08	3
Q17	0	0,559	1,2	3
Q18	0	0,606	0,908	3
Q19	0	0,259	0,911	3
Q20	0	0,797	0,949	3
Q21	0	2,029	2,212	3
Q22	0	0,124	0,989	3
Q23	0	0,447	1,284	3
Q24	0	0,168	1,053	3
Q25	0	0,921	1,292	3
Q26	0	0,265	0,73	3
Q27	0	0,264	1,341	3
Q28	0	0,765	2,082	3
Q29	0	0,697	1,99	3
Q30	0	0,505	1,978	3

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Tabela 2 - Parâmetros de calibração segundo o MRG (continua)

Questões	a	b_1	b_2	b_3
Q1	1,646538	-3,07695	-2,30444	-1,98809
Q2	1,760583	-2,54637	-1,98798	-1,54891
Q3	1,291627	-2,53421	-1,29048	-0,88873
Q4	1,281235	-1,97491	-1,33371	-0,62274
Q5	1,161561	-3,03873	-2,61666	-1,34634
Q6	0,619632	-6,46244	-1,22143	1,424341
Q7	0,914791	-2,03573	-0,85995	-0,1718
Q8	0,846897	-2,96379	-1,54135	-0,35126
Q9	1,62699	-2,38238	-1,71302	-1,30086
Q10	1,380007	-2,51363	-1,7278	-0,9082
Q11	1,233053	-2,36359	-1,89471	-1,23837
Q12	2,134039	-2,3749	-1,90795	-1,63113
Q13	1,385339	-2,78518	-1,91657	-0,88744
Q14	1,547535	-3,04936	-2,26182	-1,6839
Q15	1,777579	-2,27173	-1,92737	-1,55046
Q16	0,976367	-2,92098	-1,25234	-0,48241
Q17	1,787695	-2,61161	-1,91861	-1,14409
Q18	1,261764	-3,30589	-2,57469	-1,82089
Q19	1,425376	-2,90112	-2,1847	-1,6696
Q20	1,3879	-2,776	-1,78324	-1,26185
Q21	0,514758	-5,09789	-0,28439	1,5426
Q22	1,40619	-2,33079	-1,60901	-1,01838
Q23	1,449578	-3,1512	-2,14459	-1,55778
Q24	0,317504	-5,72157	-1,49374	0,381721
Q25	0,937704	-3,47626	-0,95391	-0,28517
Q26	1,568135	-2,59218	-1,97688	-1,54152
Q27	1,682288	-3,49917	-2,97949	-2,53933
Q28	0,798272	-4,05431	-2,80926	-1,16083
Q29	1,009394	-2,92136	-1,3493	0,302871
Q30	0,719784	-3,79112	-2,07411	-0,93571

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Tabela 1 - Parâmetros de calibração segundo o MRN (continua)

Questões	a_1	a_2	a_3	a_4
Q31	0	0,271	0,767	3
Q32	0	0,998	1,003	3
Q33	0	0,405	1,385	3
Q34	0	0,269	0,984	3
Q35	0	0,478	1,004	3
Q36	ANULADA			
Q37	0	0,1	0,7	3
Q38	0	0,297	0,709	3
Q39	0	0,85	1,403	3
Q40	0	0,034	0,387	3
Q41	0	0,981	1,428	3
Q42	0	1,138	1,851	3
Q43	0	0,408	0,503	3
Q44	0	0,588	0,773	3
Q45	0	0,526	1,562	3
Q46	0	0,14	0,507	3
Q47	ANULADA			
Q48	0	0,18	0,76	3
Q49	0	0,75	2,055	3
Q50	0	0,483	2,036	3
Q51	0	0,183	1,061	3
Q52	0	0,174	0,257	3
Q53	0	0,352	0,364	3
Q54	0	0,49	0,85	3
Q55	0	0,12	0,984	3
Q56	0	0,979	1,963	3
Q57	0	1,319	2,056	3
Q58	0	0,419	1,363	3
Q59	0	1,166	2,619	3
Q60	0	0,757	2,058	3
Q61	0	0	0,358	3

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Tabela 2 - Parâmetros de calibração segundo o MRG (continua)

Questões	a	b_1	b_2	b_3
Q31	2,251546	-2,65393	-2,20036	-1,91467
Q32	0,960697	-3,06751	-1,42189	-0,53853
Q33	0,823871	-2,46525	-1,0448	0,792644
Q34	2,028022	-2,3523	-1,91453	-1,2347
Q35	2,308643	-2,50544	-2,12374	-1,38146
Q36	ANULADA			
Q37	1,332777	-2,6987	-1,87332	-1,04477
Q38	1,503931	-1,59117	-0,97503	-0,49212
Q39	1,993625	-2,59669	-1,88844	-1,54516
Q40	1,829211	-2,84447	-2,39156	-2,05211
Q41	1,262879	-2,78351	-1,9561	-1,08148
Q42	0,819033	-3,06085	-1,7408	0,072026
Q43	1,550613	-2,40948	-1,64442	-1,1854
Q44	1,492872	-2,90253	-2,03912	-1,6085
Q45	1,505165	-2,86503	-2,42441	-1,69657
Q46	2,231841	-1,99822	-1,67847	-1,44052
Q47	ANULADA			
Q48	1,340361	-2,06915	-1,33694	-0,74569
Q49	0,915338	-2,20753	-1,30843	0,384122
Q50	0,954358	-3,00731	-2,15216	-0,44693
Q51	0,669452	-3,13613	-0,84559	1,175562
Q52	1,20064	-2,09352	-1,43468	-0,86123
Q53	1,431861	-3,02253	-2,16812	-1,59336
Q54	0,424604	-3,4151	-2,29986	2,048274
Q55	1,542431	-2,28849	-1,80174	-0,65999
Q56	1,490097	-2,39051	-1,07347	-0,51184
Q57	0,860042	-4,04363	-3,22672	-2,05102
Q58	1,548218	-2,51762	-1,61718	-0,66315
Q59	0,529047	-5,01688	-2,85688	0,814603
Q60	0,521127	-5,57034	-3,08796	-0,65554
Q61	1,963866	-2,17194	-1,79294	-1,46702

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Tabela 1 - Parâmetros de calibração segundo o MRN (conclusão)

Questões	a_1	a_2	a_3	a_4
Q62	0	1,305	2,393	3
Q63	0	0,416	0,54	3
Q64	0	0,6	0,674	3
Q65	0	0,646	1,007	3
Q66	0	0,68	0,735	3
Q67	0	0,546	1,603	3
Q68	0	0,11	0,755	3
Q69	0	0,42	2,109	3
Q70	0	0,641	2,315	3
Q71	0	1,002	1,211	3
Q72	0	0,747	1,227	3
Q73	0	0,979	1,36	3
Q74	0	0,561	0,687	3
Q75	0	0,952	1,651	3
Q76	0	0,257	0,531	3
Q77	0	0,109	0,541	3
Q78	0	0,363	1,431	3
Q79	0	0,292	0,892	3
Q80	0	1,008	2,013	3
Q81	0	0,225	0,713	3
Q82	0	0,615	1,433	3
Q83	0	0,317	1,346	3
Q84	0	0,042	0,783	3
Q85	0	0,48	0,741	3
Q86	0	0,422	0,524	3
Q87	0	1,285	1,372	3
Q88	ANULADA			
Q89	0	0,011	0,197	3
Q90	0	0,285	0,379	3
Q91	0	0,616	1,911	3

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Tabela 2 - Parâmetros de calibração segundo o MRG (conclusão)

Questões	a	b_1	b_2	b_3
Q62	0,482925	-5,81267	-1,68578	0,789758
Q63	1,9747	-2,54042	-2,17943	-1,68905
Q64	1,150708	-2,35785	-1,46993	-0,2865
Q65	2,024831	-2,49733	-2,07326	-1,66384
Q66	1,299663	-3,13954	-2,21345	-1,66605
Q67	1,745176	-2,89912	-2,40077	-1,89817
Q68	0,645528	-4,03815	-2,05855	-1,00175
Q69	0,747278	-2,64658	-1,14671	0,221786
Q70	0,879001	-4,36064	-3,27993	-0,57702
Q71	0,707123	-3,70874	-0,04403	1,157273
Q72	0,969134	-3,81397	-2,4249	-1,40508
Q73	0,956969	-3,56287	-1,74977	-0,56063
Q74	1,003806	-3,13259	-2,052	-1,24308
Q75	1,328491	-2,97447	-2,22648	-0,84665
Q76	2,069124	-2,81485	-2,35677	-2,04552
Q77	0,971523	-2,80088	-1,67108	-0,53724
Q78	1,211327	-3,00807	-2,23686	-1,41934
Q79	1,685882	-2,88971	-2,30549	-1,90317
Q80	0,742366	-3,94632	-1,53506	-0,30941
Q81	0,914699	-3,19842	-1,98895	-1,26305
Q82	0,969691	-3,06155	-1,84963	-1,0854
Q83	1,547335	-2,40789	-1,75902	-0,87419
Q84	0,652644	-3,20654	-1,68082	0,098985
Q85	1,183209	-2,57622	-1,62796	-0,7395
Q86	1,891406	-2,34823	-1,83944	-1,60213
Q87	0,855663	-3,57283	-1,07922	-0,5217
Q88	ANULADA			
Q89	1,795803	-2,09106	-1,43294	-1,01728
Q90	1,529504	-2,56419	-1,91268	-1,55942
Q91	0,970969	-4,08051	-3,12513	-1,50111

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Essas tabelas mostram os resultados das análises conduzidas na presente pesquisa no *corpus* selecionado, composto pela base de dados com as respostas dos estudantes do 5º ano aos itens de língua portuguesa da Prova São Paulo 2018, à luz dos modelos de resposta nominal e de respostas graduais de Samejima da TRI.

No próximo passo das análises, o da interpretação pedagógica dos itens, foram utilizados os 88 itens de língua portuguesa aplicados aos estudantes do 5º ano da rede municipal de ensino de São Paulo. Cabe esclarecer que esses itens fazem parte do acervo sigiloso do banco de itens da Secretaria Municipal de Ensino de São Paulo (SME-SP) e que o acesso a esse material se deu mediante autorização do referido órgão (Anexo).

A metodologia adotada para a interpretação dos itens é baseada na que é utilizada pelo Inep atualmente nas escalas de proficiência do Saeb (INEP, 2018, p. 47), segundo a qual

> [...] os itens são descritos pedagogicamente de acordo com três elementos estruturais: 1) Operação cognitiva: traduz as ações requeridas ao participante do teste para resolver a situação-problema proposta pelo item; 2) Objeto do conhecimento: refere-se aos conhecimentos escolares solicitados ou mobilizados no item para que o respondente execute a operação cognitiva visando à resolução do item; 3) Contexto: considera as situações envolvidas no problema construído pelo item.

Considerando o enfoque da escala proposta neste estudo há, porém, duas diferenças significativas observadas. A primeira é em relação à operação cognitiva que, na presente pesquisa, ao invés de dar enfoque à ação requerida do estudante para resolver a situação-problema, se buscou traduzir a ação realizada que configurou o erro cometido. Essa ação corresponde ao estágio de desenvolvimento da habilidade alcançado pelo aluno, entendendo, assim como Oliveira, Franco e Soares (2007), que as habilidades existem em um *continuum*, que vai desde o seu início de desenvolvimento, passando pelo seu processamento até chegar à sua consolidação.

A segunda diz respeito aos conhecimentos escolares mobilizados pelo estudante. Na presente proposta de escala, os conhecimentos descritos são aqueles que o estudante demonstrou ter apreendido na etapa de desenvolvimento da habilidade em que se encontra.

Seguindo essa metodologia, foi possível construir a escala interpretada que tem como base os erros cometidos pelos estudantes durante a realização do teste de língua portuguesa da Prova São Paulo. Tal escala é apresentada na próxima seção deste estudo.

Resultados e discussão

1. Definição da escala

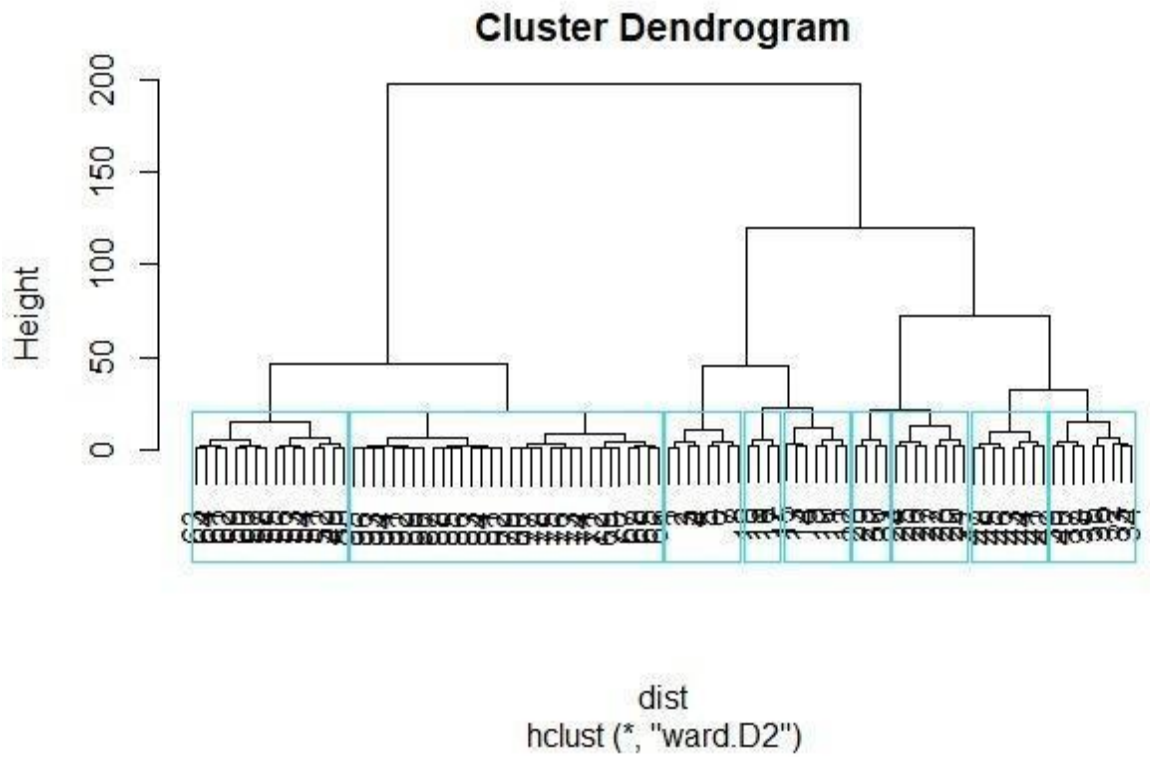
Com base nos resultados da calibração dos itens segundo o MRN e o MRG, se partiu para a definição da escala. Novamente, optou-se por utilizar a metodologia aplicada por Gabriela Thamara de Freitas Barros (2006, p. 83), uma vez que, conforme a autora, não se encontra na literatura opções claras sobre como dividir e interpretar escalas.

Começando pela seleção dos pontos para a interpretação da escala que aqui se propõe, destaca-se que, em seu trabalho, Barros (2006) utilizou o Método de Ward para definir agrupamentos hierárquicos. Segundo Hair (2009, p. 430), essa técnica busca "agregar objetos com base nas características que eles possuem". Desse modo, um indivíduo é muito semelhante aos demais indivíduos do mesmo grupo, mas diferente dos indivíduos dos outros grupos.

As variáveis selecionadas para essa análise foram as habilidades dos indivíduos e as probabilidades de resposta por questão. Os valores utilizados para as habilidades variam de 0 a 100, com intervalos de 10 unidades[11]. Com base nas probabilidades e utilizando a técnica de cluster, foram feitos os cortes na escala. Tal procedimento foi realizado por meio do *software* livre R de análise de dados. A Figura 12 a seguir mostra o gráfico de dendrograma que definiu esses cortes, em que os valores de habilidade são inseridos no eixo vertical e o número de agrupamentos no eixo horizontal.

[11] Não foi possível utilizar nesta pesquisa a mesma escala adotada para a apresentação dos resultados da Prova São Paulo em virtude da diferença nos modelos da TRI selecionados para a realização da análise conduzida na presente pesquisa.

Figura 12 - Dendrograma



Seguindo o procedimento adotado por Barros (2006), após a definição das faixas de habilidade por grupo, foram calculadas as médias das probabilidades referentes a cada opção, por grupo definido, buscando o valor em que a probabilidade era mais alta dentre os valores calculados. Ademais, se buscou verificar até que faixa de habilidade cada opção de resposta se destaca em relação às demais. O resultado dessa análise é apresentado na Tabela 3, em que são mostrados somente os valores calculados para os distratores, ou seja, as opções erradas das questões, que são o foco de estudo da presente pesquisa.

Ressalta-se, na leitura desta tabela, que os grupos englobam os valores de habilidade dos indivíduos. O grupo 1, por exemplo, reúne aqueles que apresentaram valores de habilidade maiores ou iguais a zero e menores que 12. O grupo 2, indivíduos com valores de habilidade maiores ou iguais a 12 e menores que 19 e assim sucessivamente. O Quadro 10 a seguir detalha esses grupos.

Quadro 10 - Definição dos grupos de proficiência

Grupos	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
Valores de proficiência	≥ 0 e < 12	≥ 12 e < 19	≥ 19 e < 23	≥ 23 e < 31	≥ 31 e < 35	≥ 35 e < 44	≥ 44 e < 52	≥ 52 e < 68	≥ 68 e < 100

Fonte: Elaborado pela autora (2022).

Para orientar a leitura dos dados da tabela, as células coloridas em verde indicam opções cuja probabilidade de marcação é maior ou igual a 0,5; as células em amarelo são aquelas com valores de probabilidade maiores que 0,3 e menores que 0,5; e as demais são as que apresentaram probabilidade de marcação inferior a 0,3.

Tabela 3 - Probabilidade média por questão/alternativa e grupos (continua)

Grupo		Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
	Opções	(0,12]	(12,19]	(19,23]	(23,31]	(31,35]	(35,44]	(44,52]	(52,68]	(68,100]
Q1_1	A	0,74	0,49	0,3	0,16	0,07	0,03	0,01	0	0
Q1_2	C	0,17	0,28	0,3	0,23	0,14	0,07	0,02	0	0
Q1_3	B	0,03	0,08	0,11	0,12	0,1	0,05	0,02	0	0
Q2_1	D	0,89	0,71	0,51	0,29	0,14	0,06	0,02	0	0
Q2_2	C	0,07	0,16	0,22	0,22	0,16	0,08	0,02	0	0
Q2_3	A	0,02	0,07	0,12	0,18	0,18	0,11	0,04	0,01	0
Q3_1	C	0,82	0,66	0,51	0,34	0,21	0,11	0,04	0,01	0
Q3_2	A	0,14	0,24	0,33	0,37	0,36	0,27	0,14	0,05	0
Q3_3	D	0,02	0,04	0,06	0,09	0,12	0,12	0,09	0,04	0
Q4_1	C	0,9	0,8	0,68	0,52	0,35	0,21	0,09	0,03	0
Q4_2	B	0,05	0,1	0,15	0,19	0,2	0,16	0,09	0,03	0
Q4_3	D	0,03	0,06	0,09	0,15	0,2	0,22	0,17	0,07	0
Q5_1	A	0,69	0,51	0,37	0,24	0,14	0,08	0,03	0,01	0
Q5_2	C	0,09	0,12	0,12	0,1	0,07	0,04	0,02	0,01	0
Q5_3	B	0,16	0,25	0,32	0,35	0,33	0,25	0,15	0,05	0
Q6_1	B	0,16	0,11	0,08	0,06	0,04	0,03	0,02	0,01	0
Q6_2	A	0,67	0,65	0,61	0,56	0,5	0,42	0,32	0,2	0,04
Q6_3	C	0,14	0,18	0,23	0,27	0,32	0,36	0,39	0,36	0,14
Q7_1	A	0,82	0,72	0,62	0,5	0,38	0,26	0,15	0,07	0,01
Q7_2	D	0,11	0,16	0,21	0,24	0,26	0,25	0,19	0,1	0,01
Q7_3	B	0,03	0,05	0,07	0,1	0,13	0,15	0,15	0,1	0,01
Q8_1	B	0,66	0,52	0,42	0,31	0,22	0,15	0,08	0,04	0
Q8_2	D	0,21	0,26	0,29	0,29	0,27	0,22	0,15	0,08	0,01
Q8_3	A	0,08	0,13	0,16	0,2	0,23	0,24	0,22	0,14	0,02

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Tabela 3 - Probabilidade média por questão/alternativa e grupos (continua)

Grupo		Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
	Opções	(0,12]	(12,19]	(19,23]	(23,31]	(31,35]	(35,44]	(44,52]	(52,68]	(68,100]
Q9_1	C	0,9	0,75	0,57	0,36	0,19	0,09	0,03	0,01	0
Q9_2	B	0,07	0,15	0,23	0,26	0,22	0,13	0,05	0,01	0
Q9_3	D	0,02	0,05	0,09	0,14	0,16	0,13	0,06	0,01	0
Q10_1	D	0,84	0,68	0,52	0,34	0,2	0,1	0,04	0,01	0
Q10_2	C	0,1	0,18	0,24	0,26	0,22	0,15	0,07	0,02	0
Q10_3	A	0,04	0,09	0,15	0,22	0,27	0,25	0,16	0,05	0
Q11_1	B	0,84	0,7	0,56	0,4	0,25	0,14	0,06	0,02	0
Q11_2	D	0,06	0,11	0,13	0,14	0,12	0,09	0,04	0,01	0
Q11_3	C	0,05	0,1	0,14	0,18	0,2	0,17	0,1	0,04	0
Q12_1	D	0,94	0,8	0,6	0,33	0,13	0,05	0,01	0	0
Q12_2	B	0,03	0,11	0,2	0,23	0,16	0,07	0,02	0	0
Q12_3	A	0,01	0,04	0,08	0,13	0,13	0,07	0,02	0	0
Q13_1	B	0,78	0,59	0,42	0,26	0,14	0,07	0,03	0,01	0
Q13_2	A	0,14	0,23	0,29	0,27	0,21	0,13	0,05	0,02	0
Q13_3	D	0,06	0,12	0,2	0,29	0,34	0,3	0,18	0,06	0
Q14_1	C	0,74	0,5	0,32	0,17	0,08	0,04	0,01	0	0
Q14_2	A	0,17	0,27	0,29	0,24	0,15	0,08	0,03	0,01	0
Q14_3	D	0,05	0,12	0,18	0,21	0,19	0,12	0,05	0,01	0
Q15_1	C	0,93	0,8	0,63	0,4	0,2	0,09	0,02	0	0
Q15_2	B	0,03	0,08	0,13	0,15	0,11	0,06	0,02	0	0
Q15_3	D	0,02	0,06	0,1	0,15	0,16	0,1	0,04	0,01	0
Q16_1	B	0,69	0,53	0,41	0,3	0,2	0,12	0,06	0,02	0
Q16_2	D	0,23	0,32	0,37	0,38	0,36	0,29	0,19	0,09	0,01
Q16_3	C	0,04	0,07	0,1	0,14	0,17	0,18	0,16	0,1	0,01

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Tabela 3 - Probabilidade média por questão/alternativa e grupos (continua)

Grupo		Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
	Opções	(0,12]	(12,19]	(19,23]	(23,31]	(31,35]	(35,44]	(44,52]	(52,68]	(68,100]
Q17_1	B	0,88	0,69	0,48	0,27	0,12	0,05	0,01	0	0
Q17_2	A	0,08	0,19	0,28	0,28	0,2	0,1	0,03	0,01	0
Q17_3	D	0,03	0,09	0,17	0,28	0,33	0,25	0,11	0,02	0
Q18_1	C	0,63	0,42	0,28	0,17	0,09	0,05	0,02	0,01	0
Q18_2	B	0,18	0,22	0,21	0,17	0,11	0,06	0,03	0,01	0
Q18_3	A	0,11	0,18	0,22	0,23	0,19	0,13	0,06	0,02	0
Q19_1	B	0,76	0,56	0,38	0,23	0,12	0,06	0,02	0	0
Q19_2	A	0,14	0,22	0,25	0,22	0,15	0,09	0,03	0,01	0
Q19_3	C	0,05	0,1	0,15	0,18	0,16	0,11	0,05	0,01	0
Q20_1	A	0,79	0,6	0,43	0,26	0,14	0,07	0,03	0,01	0
Q20_2	B	0,15	0,26	0,32	0,32	0,25	0,16	0,07	0,02	0
Q20_3	D	0,03	0,07	0,11	0,16	0,18	0,15	0,08	0,03	0
Q21_1	C	0,33	0,26	0,21	0,17	0,13	0,1	0,07	0,04	0,01
Q21_2	B	0,52	0,55	0,55	0,54	0,51	0,47	0,41	0,3	0,1
Q21_3	D	0,08	0,11	0,13	0,15	0,18	0,2	0,22	0,23	0,13
Q22_1	A	0,87	0,73	0,58	0,4	0,23	0,12	0,05	0,01	0
Q22_2	C	0,08	0,15	0,21	0,24	0,22	0,15	0,07	0,02	0
Q22_3	B	0,03	0,06	0,11	0,16	0,2	0,19	0,11	0,04	0
Q23_1	C	0,7	0,47	0,3	0,17	0,08	0,04	0,01	0	0
Q23_2	A	0,21	0,32	0,35	0,29	0,2	0,11	0,04	0,01	0
Q23_3	D	0,05	0,11	0,16	0,2	0,2	0,14	0,06	0,02	0
Q24_1	B	0,35	0,3	0,27	0,24	0,21	0,18	0,15	0,11	0,05
Q24_2	C	0,32	0,32	0,32	0,31	0,29	0,27	0,25	0,21	0,11
Q24_3	A	0,12	0,13	0,13	0,14	0,14	0,15	0,15	0,14	0,1

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Tabela 3 - Probabilidade média por questão/alternativa e grupos (continua)

Grupo		Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
	Opções	(0,12]	(12,19]	(19,23]	(23,31]	(31,35]	(35,44]	(44,52]	(52,68]	(68,100]
Q25_1	D	0,56	0,4	0,3	0,21	0,13	0,08	0,04	0,02	0
Q25_2	A	0,37	0,47	0,52	0,52	0,49	0,4	0,28	0,13	0,01
Q25_3	C	0,03	0,05	0,08	0,1	0,13	0,15	0,15	0,1	0,01
Q26_1	C	0,85	0,67	0,49	0,3	0,15	0,07	0,02	0	0
Q26_2	A	0,09	0,17	0,23	0,22	0,17	0,09	0,03	0,01	0
Q26_3	D	0,03	0,07	0,12	0,16	0,16	0,11	0,05	0,01	0
Q27_1	C	0,59	0,33	0,17	0,08	0,03	0,01	0	0	0
Q27_2	B	0,18	0,21	0,16	0,09	0,04	0,02	0,01	0	0
Q27_3	D	0,1	0,17	0,18	0,13	0,07	0,03	0,01	0	0
Q28_1	B	0,44	0,31	0,23	0,17	0,11	0,08	0,04	0,02	0
Q28_2	C	0,24	0,24	0,22	0,18	0,14	0,1	0,06	0,03	0
Q28_3	A	0,21	0,27	0,3	0,32	0,31	0,27	0,2	0,11	0,01
Q29_1	B	0,69	0,53	0,41	0,29	0,19	0,12	0,06	0,02	0
Q29_2	D	0,22	0,31	0,36	0,37	0,34	0,27	0,17	0,07	0
Q29_3	A	0,07	0,12	0,17	0,25	0,32	0,38	0,38	0,25	0,02
Q30_1	C	0,49	0,37	0,29	0,22	0,16	0,11	0,07	0,03	0
Q30_2	A	0,28	0,3	0,29	0,27	0,24	0,19	0,13	0,07	0,01
Q30_3	D	0,11	0,15	0,18	0,19	0,2	0,19	0,16	0,1	0,02
Q31_1	A	0,91	0,71	0,45	0,21	0,07	0,02	0	0	0
Q31_2	C	0,05	0,16	0,24	0,2	0,1	0,04	0,01	0	0
Q31_3	B	0,02	0,06	0,12	0,15	0,11	0,04	0,01	0	0
Q32_1	C	0,65	0,5	0,38	0,27	0,18	0,11	0,06	0,02	0
Q32_2	A	0,25	0,33	0,37	0,37	0,34	0,27	0,17	0,08	0,01
Q32_3	D	0,05	0,09	0,13	0,17	0,2	0,21	0,18	0,1	0,01

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Tabela 3 - Probabilidade média por questão/alternativa e grupos (continua)

Grupo		Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
	Opções	(0,12]	(12,19]	(19,23]	(23,31]	(31,35]	(35,44]	(44,52]	(52,68]	(68,100]
Q33_1	B	0,74	0,62	0,52	0,41	0,31	0,22	0,13	0,06	0,01
Q33_2	A	0,16	0,22	0,26	0,28	0,28	0,25	0,19	0,11	0,01
Q33_3	C	0,08	0,12	0,16	0,22	0,28	0,33	0,36	0,3	0,06
Q34_1	A	0,94	0,8	0,6	0,35	0,15	0,05	0,01	0	0
Q34_2	C	0,04	0,1	0,18	0,21	0,15	0,07	0,02	0	0
Q34_3	B	0,02	0,07	0,15	0,27	0,33	0,22	0,07	0,01	0
Q35_1	B	0,94	0,77	0,53	0,26	0,09	0,03	0	0	0
Q35_2	C	0,03	0,12	0,2	0,19	0,1	0,04	0,01	0	0
Q35_3	A	0,02	0,09	0,21	0,36	0,37	0,2	0,05	0,01	0
Q37_1	D	0,8	0,62	0,46	0,29	0,17	0,09	0,03	0,01	0
Q37_2	A	0,13	0,21	0,26	0,26	0,21	0,13	0,06	0,02	0
Q37_3	B	0,05	0,11	0,17	0,23	0,27	0,23	0,14	0,05	0
Q38_1	A	0,96	0,9	0,81	0,65	0,46	0,27	0,1	0,03	0
Q38_2	B	0,02	0,06	0,1	0,17	0,22	0,21	0,12	0,04	0
Q38_3	C	0,01	0,02	0,04	0,08	0,13	0,17	0,15	0,06	0
Q39_1	C	0,9	0,71	0,48	0,25	0,1	0,04	0,01	0	0
Q39_2	B	0,07	0,2	0,31	0,31	0,21	0,09	0,02	0	0
Q39_3	D	0,01	0,04	0,09	0,15	0,16	0,1	0,03	0	0
Q40_1	B	0,83	0,59	0,38	0,19	0,08	0,03	0,01	0	0
Q40_2	C	0,09	0,17	0,2	0,15	0,08	0,04	0,01	0	0
Q40_3	A	0,04	0,09	0,14	0,14	0,1	0,05	0,01	0	0
Q41_1	C	0,77	0,59	0,43	0,28	0,16	0,09	0,04	0,01	0
Q41_2	D	0,14	0,21	0,25	0,24	0,19	0,13	0,06	0,02	0
Q41_3	A	0,06	0,12	0,18	0,24	0,27	0,23	0,14	0,05	0

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Tabela 3 - Probabilidade média por questão/alternativa e grupos (continua)

Grupo		Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
	Opções	(0,12]	(12,19]	(19,23]	(23,31]	(31,35]	(35,44]	(44,52]	(52,68]	(68,100]
Q42_1	B	0,63	0,5	0,4	0,3	0,22	0,15	0,08	0,04	0
Q42_2	D	0,2	0,25	0,26	0,26	0,23	0,19	0,13	0,07	0,01
Q42_3	A	0,12	0,18	0,23	0,29	0,33	0,35	0,33	0,23	0,04
Q43_1	B	0,88	0,73	0,56	0,36	0,19	0,09	0,03	0,01	0
Q43_2	A	0,08	0,17	0,25	0,28	0,25	0,15	0,06	0,01	0
Q43_3	C	0,02	0,05	0,09	0,14	0,17	0,15	0,08	0,02	0
Q44_1	B	0,77	0,56	0,38	0,22	0,11	0,05	0,02	0	0
Q44_2	A	0,15	0,26	0,31	0,28	0,2	0,11	0,04	0,01	0
Q44_3	D	0,03	0,08	0,12	0,15	0,15	0,1	0,05	0,01	0
Q45_1	A	0,78	0,57	0,39	0,22	0,11	0,05	0,02	0	0
Q45_2	C	0,09	0,15	0,16	0,13	0,08	0,04	0,02	0	0
Q45_3	B	0,08	0,16	0,23	0,26	0,22	0,14	0,06	0,01	0
Q46_1	A	0,98	0,91	0,78	0,51	0,24	0,09	0,02	0	0
Q46_2	B	0,01	0,04	0,1	0,16	0,15	0,07	0,02	0	0
Q46_3	D	0	0,02	0,05	0,1	0,13	0,08	0,02	0	0
Q48_1	B	0,9	0,79	0,66	0,49	0,31	0,18	0,07	0,02	0
Q48_2	A	0,06	0,12	0,18	0,23	0,23	0,18	0,1	0,03	0
Q48_3	C	0,02	0,05	0,08	0,13	0,18	0,19	0,14	0,05	0
Q49_1	D	0,8	0,68	0,58	0,46	0,34	0,23	0,13	0,06	0
Q49_2	B	0,1	0,15	0,18	0,2	0,2	0,17	0,12	0,06	0,01
Q49_3	A	0,08	0,13	0,18	0,24	0,31	0,35	0,36	0,26	0,03
Q50_1	B	0,67	0,51	0,4	0,28	0,19	0,12	0,06	0,02	0
Q50_2	D	0,15	0,19	0,2	0,19	0,16	0,11	0,07	0,03	0
Q50_3	C	0,14	0,22	0,29	0,35	0,38	0,37	0,3	0,16	0,01

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Tabela 3 - Probabilidade média por questão/alternativa e grupos (continua)

Grupo		Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
	Opções	(0,12]	(12,19]	(19,23]	(23,31]	(31,35]	(35,44]	(44,52]	(52,68]	(68,100]
Q51_1	C	0,6	0,49	0,41	0,32	0,25	0,18	0,12	0,06	0,01
Q51_2	D	0,27	0,33	0,35	0,36	0,36	0,33	0,26	0,17	0,03
Q51_3	A	0,09	0,13	0,16	0,21	0,25	0,29	0,32	0,3	0,11
Q52_1	B	0,88	0,76	0,64	0,48	0,33	0,2	0,09	0,03	0
Q52_2	A	0,06	0,11	0,16	0,19	0,19	0,15	0,09	0,03	0
Q52_3	C	0,03	0,06	0,09	0,13	0,16	0,16	0,12	0,05	0
Q53_1	C	0,73	0,51	0,34	0,2	0,1	0,05	0,02	0	0
Q53_2	B	0,17	0,27	0,3	0,25	0,17	0,1	0,04	0,01	0
Q53_3	A	0,05	0,11	0,16	0,2	0,19	0,13	0,06	0,02	0
Q54_1	A	0,53	0,46	0,41	0,36	0,31	0,26	0,2	0,14	0,04
Q54_2	C	0,11	0,12	0,12	0,11	0,11	0,1	0,09	0,07	0,03
Q54_3	B	0,27	0,32	0,35	0,38	0,4	0,42	0,43	0,41	0,25
Q55_1	D	0,9	0,76	0,6	0,4	0,23	0,11	0,04	0,01	0
Q55_2	C	0,05	0,11	0,16	0,18	0,16	0,1	0,04	0,01	0
Q55_3	B	0,04	0,1	0,19	0,3	0,4	0,38	0,24	0,07	0
Q56_1	C	0,88	0,73	0,56	0,37	0,21	0,1	0,03	0,01	0
Q56_2	A	0,1	0,22	0,34	0,43	0,44	0,34	0,17	0,05	0
Q56_3	D	0,01	0,03	0,05	0,1	0,16	0,2	0,16	0,06	0
Q57_1	B	0,43	0,3	0,22	0,15	0,1	0,06	0,03	0,01	0
Q57_2	A	0,17	0,16	0,14	0,11	0,08	0,06	0,03	0,01	0
Q57_3	D	0,2	0,24	0,25	0,23	0,2	0,15	0,1	0,05	0
Q58_1	B	0,86	0,7	0,52	0,32	0,17	0,08	0,03	0,01	0
Q58_2	A	0,1	0,21	0,29	0,33	0,28	0,17	0,07	0,02	0
Q58_3	C	0,03	0,07	0,14	0,24	0,33	0,33	0,22	0,07	0

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Tabela 3 - Probabilidade média por questão/alternativa e grupos (continua)

Grupo		Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
	Opções	(0,12]	(12,19]	(19,23]	(23,31]	(31,35]	(35,44]	(44,52]	(52,68]	(68,100]
Q59_1	B	0,34	0,26	0,21	0,17	0,13	0,1	0,07	0,04	0,01
Q59_2	D	0,28	0,26	0,25	0,22	0,19	0,16	0,12	0,08	0,02
Q59_3	C	0,3	0,36	0,4	0,43	0,45	0,45	0,43	0,36	0,14
Q60_1	B	0,28	0,21	0,17	0,14	0,11	0,08	0,06	0,03	0,01
Q60_2	D	0,31	0,28	0,26	0,23	0,2	0,16	0,12	0,08	0,02
Q60_3	A	0,25	0,28	0,3	0,31	0,3	0,29	0,26	0,19	0,06
Q61_1	A	0,95	0,85	0,68	0,43	0,21	0,08	0,02	0	0
Q61_2	D	0,02	0,07	0,14	0,18	0,15	0,07	0,02	0	0
Q61_3	C	0,01	0,04	0,08	0,13	0,15	0,1	0,03	0,01	0
Q62_1	D	0,27	0,21	0,17	0,14	0,11	0,09	0,06	0,04	0,01
Q62_2	A	0,46	0,45	0,43	0,4	0,37	0,32	0,26	0,18	0,06
Q62_3	B	0,17	0,21	0,23	0,25	0,27	0,29	0,29	0,26	0,12
Q63_1	B	0,91	0,73	0,51	0,27	0,11	0,04	0,01	0	0
Q63_2	C	0,04	0,11	0,17	0,15	0,09	0,04	0,01	0	0
Q63_3	D	0,03	0,09	0,17	0,23	0,2	0,1	0,03	0	0
Q64_1	B	0,83	0,69	0,56	0,41	0,27	0,16	0,07	0,02	0
Q64_2	C	0,1	0,17	0,22	0,25	0,24	0,18	0,11	0,04	0
Q64_3	A	0,05	0,1	0,15	0,23	0,29	0,32	0,28	0,14	0,01
Q65_1	A	0,92	0,75	0,53	0,29	0,12	0,04	0,01	0	0
Q65_2	D	0,04	0,12	0,19	0,19	0,12	0,05	0,01	0	0
Q65_3	B	0,02	0,07	0,13	0,19	0,18	0,09	0,03	0	0
Q66_1	B	0,68	0,47	0,32	0,19	0,1	0,05	0,02	0,01	0
Q66_2	D	0,19	0,27	0,29	0,25	0,18	0,1	0,04	0,01	0
Q66_3	C	0,06	0,11	0,15	0,17	0,16	0,12	0,06	0,02	0

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Tabela 3 - Probabilidade média por questão/alternativa e grupos (continua)

Grupo		Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
	Opções	(0,12]	(12,19]	(19,23]	(23,31]	(31,35]	(35,44]	(44,52]	(52,68]	(68,100]
Q67_1	A	0,81	0,57	0,36	0,18	0,08	0,03	0,01	0	0
Q67_2	C	0,1	0,19	0,21	0,16	0,09	0,04	0,01	0	0
Q67_3	B	0,05	0,12	0,19	0,21	0,16	0,08	0,03	0,01	0
Q68_1	D	0,45	0,35	0,28	0,22	0,16	0,12	0,08	0,04	0,01
Q68_2	A	0,29	0,31	0,3	0,28	0,25	0,2	0,15	0,09	0,02
Q68_3	B	0,11	0,13	0,15	0,16	0,17	0,16	0,14	0,1	0,02
Q69_1	B	0,69	0,58	0,48	0,39	0,3	0,21	0,13	0,07	0,01
Q69_2	A	0,18	0,23	0,26	0,27	0,27	0,24	0,19	0,11	0,02
Q69_3	D	0,08	0,11	0,15	0,18	0,22	0,24	0,25	0,19	0,04
Q70_1	A	0,37	0,24	0,17	0,12	0,07	0,05	0,02	0,01	0
Q70_2	B	0,23	0,21	0,18	0,14	0,1	0,07	0,04	0,02	0
Q70_3	C	0,34	0,45	0,5	0,53	0,52	0,46	0,34	0,19	0,02
Q71_1	D	0,5	0,39	0,31	0,23	0,17	0,12	0,07	0,04	0,01
Q71_2	A	0,43	0,51	0,55	0,57	0,56	0,52	0,44	0,3	0,06
Q71_3	B	0,04	0,06	0,08	0,1	0,13	0,16	0,2	0,2	0,07
Q72_1	C	0,48	0,33	0,23	0,15	0,1	0,06	0,03	0,01	0
Q72_2	D	0,3	0,32	0,3	0,25	0,19	0,13	0,07	0,03	0
Q72_3	B	0,13	0,18	0,22	0,24	0,23	0,19	0,13	0,06	0
Q73_1	A	0,54	0,38	0,28	0,19	0,12	0,07	0,04	0,01	0
Q73_2	D	0,33	0,39	0,41	0,38	0,32	0,24	0,14	0,06	0
Q73_3	C	0,09	0,14	0,19	0,24	0,27	0,27	0,22	0,12	0,01
Q74_1	D	0,65	0,48	0,36	0,25	0,16	0,1	0,05	0,02	0
Q74_2	C	0,2	0,25	0,26	0,24	0,2	0,14	0,08	0,03	0
Q74_3	B	0,08	0,13	0,16	0,19	0,2	0,17	0,12	0,06	0

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Tabela 3 - Probabilidade média por questão/alternativa e grupos (continua)

Grupo		Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
	Opções	(0,12]	(12,19]	(19,23]	(23,31]	(31,35]	(35,44]	(44,52]	(52,68]	(68,100]
Q75_1	C	0,73	0,53	0,37	0,22	0,12	0,06	0,02	0,01	0
Q75_2	B	0,15	0,22	0,24	0,21	0,15	0,09	0,04	0,01	0
Q75_3	A	0,1	0,2	0,3	0,39	0,43	0,37	0,22	0,08	0
Q76_1	D	0,86	0,62	0,37	0,17	0,06	0,02	0	0	0
Q76_2	C	0,08	0,18	0,23	0,17	0,08	0,03	0,01	0	0
Q76_3	B	0,03	0,08	0,14	0,15	0,1	0,04	0,01	0	0
Q77_1	C	0,71	0,56	0,44	0,32	0,22	0,14	0,07	0,03	0
Q77_2	D	0,17	0,23	0,26	0,26	0,24	0,18	0,11	0,05	0
Q77_3	A	0,08	0,13	0,17	0,22	0,26	0,26	0,22	0,12	0,01
Q78_1	B	0,7	0,52	0,37	0,24	0,14	0,07	0,03	0,01	0
Q78_2	A	0,15	0,21	0,23	0,2	0,15	0,09	0,04	0,01	0
Q78_3	D	0,08	0,15	0,2	0,24	0,23	0,18	0,1	0,04	0
Q79_1	B	0,8	0,57	0,37	0,19	0,09	0,04	0,01	0	0
Q79_2	A	0,11	0,21	0,24	0,19	0,12	0,05	0,02	0	0
Q79_3	C	0,04	0,1	0,14	0,16	0,13	0,07	0,02	0,01	0
Q80_1	B	0,46	0,34	0,26	0,19	0,14	0,1	0,06	0,03	0
Q80_2	A	0,37	0,41	0,42	0,39	0,35	0,29	0,21	0,11	0,02
Q80_3	C	0,09	0,13	0,16	0,19	0,21	0,22	0,21	0,15	0,03
Q81_1	A	0,62	0,47	0,36	0,26	0,17	0,11	0,06	0,02	0
Q81_2	B	0,21	0,26	0,27	0,25	0,21	0,16	0,1	0,04	0
Q81_3	D	0,07	0,11	0,14	0,16	0,16	0,15	0,11	0,06	0
Q82_1	B	0,66	0,5	0,38	0,27	0,18	0,11	0,06	0,02	0
Q82_2	D	0,2	0,26	0,28	0,27	0,23	0,18	0,11	0,04	0
Q82_3	C	0,07	0,11	0,14	0,17	0,18	0,17	0,13	0,06	0

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Tabela 3 - Probabilidade média por questão/alternativa e grupos (conclusão)

Grupo		Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
	Opções	(0,12]	(12,19]	(19,23]	(23,31]	(31,35]	(35,44]	(44,52]	(52,68]	(68,100]
Q83_1	A	0,88	0,73	0,56	0,36	0,19	0,09	0,03	0,01	0
Q83_2	C	0,07	0,15	0,22	0,24	0,2	0,12	0,05	0,01	0
Q83_3	B	0,03	0,09	0,16	0,25	0,32	0,29	0,17	0,05	0
Q84_1	B	0,58	0,48	0,4	0,32	0,25	0,18	0,12	0,06	0,01
Q84_2	A	0,21	0,23	0,24	0,24	0,22	0,19	0,15	0,09	0,02
Q84_3	C	0,13	0,18	0,21	0,24	0,27	0,28	0,27	0,21	0,06
Q85_1	B	0,79	0,64	0,5	0,34	0,22	0,12	0,05	0,02	0
Q85_2	C	0,13	0,2	0,25	0,27	0,24	0,18	0,09	0,03	0
Q85_3	D	0,05	0,1	0,14	0,2	0,25	0,25	0,18	0,08	0
Q86_1	A	0,93	0,79	0,6	0,36	0,17	0,07	0,02	0	0
Q86_2	C	0,04	0,12	0,2	0,23	0,17	0,09	0,02	0	0
Q86_3	B	0,01	0,03	0,06	0,1	0,11	0,07	0,02	0	0
Q87_1	B	0,53	0,39	0,3	0,21	0,14	0,09	0,05	0,02	0
Q87_2	D	0,37	0,45	0,48	0,48	0,44	0,37	0,26	0,13	0,01
Q87_3	C	0,03	0,05	0,07	0,09	0,11	0,12	0,11	0,07	0,01
Q89_1	B	0,95	0,85	0,7	0,47	0,25	0,11	0,03	0,01	0
Q89_2	C	0,04	0,1	0,18	0,27	0,27	0,17	0,06	0,01	0
Q89_3	A	0,01	0,03	0,06	0,12	0,17	0,16	0,08	0,02	0
Q90_1	B	0,85	0,68	0,5	0,31	0,16	0,08	0,02	0,01	0
Q90_2	C	0,09	0,17	0,23	0,23	0,18	0,1	0,04	0,01	0
Q90_3	A	0,02	0,06	0,09	0,13	0,13	0,09	0,04	0,01	0
Q91_1	C	0,42	0,27	0,19	0,12	0,07	0,04	0,02	0,01	0
Q91_2	D	0,22	0,21	0,18	0,14	0,09	0,06	0,03	0,01	0
Q91_3	B	0,25	0,33	0,37	0,37	0,33	0,25	0,16	0,07	0

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Observa-se, de uma forma geral, que para os distratores que apresentaram probabilidade de marcação maior que 0,5, essa probabilidade tende a aumentar quanto menor for a proficiência do respondente e diminuir quanto maior for essa proficiência. Essa constatação parece indicar que, da mesma forma que a escala de proficiência tradicional[12], construída com base nos acertos dos respondentes, parece existir uma cumulatividade entre os grupos. Isso quer dizer que um erro posicionado no grupo 4, por exemplo, provavelmente, também é cometido pelos respondentes dos grupos inferiores a esse (3, 2 e 1).

Há, no entanto, uma particularidade interessante quanto a essa "escala do erro". Nas questões 25 e 71, se observa que um distrator apresenta probabilidade maior de marcação em determinados grupos, enquanto outro, em outros grupos. A Tabela 4 abaixo ilustra esses casos.

Tabela 4 - Probabilidade média por questão/alternativa e grupos (continua)

Grupos		**Grupo 1**	**Grupo 2**	**Grupo 3**	**Grupo 4**	**Grupo 5**	**Grupo 6**	**Grupo 7**	**Grupo 8**	**Grupo 9**
Opções		(0,12]	(12,19]	(19,23]	(23,31]	(31,35]	(35,44]	(44,52]	(52,68]	(68,100]
Q25_1	D	0,56	0,4	0,3	0,21	0,13	0,08	0,04	0,02	0
Q25_2	A	0,37	0,47	0,52	0,52	0,49	0,4	0,28	0,13	0,01
Q25_3	C	0,03	0,05	0,08	0,1	0,13	0,15	0,15	0,1	0,01
Q71_1	D	0,5	0,39	0,31	0,23	0,17	0,12	0,07	0,04	0,01
Q71_2	A	0,43	0,51	0,55	0,57	0,56	0,52	0,44	0,3	0,06
Q71_3	B	0,04	0,06	0,08	0,1	0,13	0,16	0,2	0,2	0,07

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Pela análise desses dados, percebe-se que, na questão 25, o distrator D é mais escolhido por indivíduos com proficiências entre 0 e 12 pontos, enquanto o distrator A é mais escolhido por indivíduos com proficiências entre 19 e 31. Na questão 71, para o grupo 1 (proficiências 0 a 12), destaca-se o distrator D e, para os grupos 2 a 6 (proficiências 12 a 44), o distrator A.

[12] Beaton e Allen explicam essa cumulatividade ao afirmar que estudantes de níveis mais altos de proficiência, geralmente, sabem e são capazes de fazer tudo o que estudantes de níveis mais baixos sabem e são capazes de fazer. Em uma escala de proficiência, isso é representado partindo-se do pressuposto de que níveis mais avançados englobam as habilidades descritas nos níveis inferiores (BEATON; ALLEN, 1992).

Tal comportamento parece ser sugestivo do que, idealmente, deveria ser a forma com que se constroem os distratores de um item: com base nos erros cometidos por indivíduos durante o processo de construção de determinada habilidade, em um *continuum*. Ou seja, indivíduos com proficiência muito baixa cometem determinado erro; indivíduos com uma proficiência maior, porém, ainda baixa, também cometem um erro, mas diferente do primeiro, e assim sucessivamente até desenvolverem a habilidade e acertarem o item.

2. Interpretação da escala

Retomando os dados apresentados na Tabela 3, como critério para o posicionamento das opções erradas, na presente pesquisa se optou por considerar apenas aquelas que apresentaram probabilidade de marcação maior que 0,5, pois se entende que esses representam os erros que a maior parte dos indivíduos de determinado grupo cometeu. Nesse cenário, 11 questões foram excluídas em virtude de apresentarem distratores com probabilidades inferiores a 0,5, a saber, as questões 24, 28, 30, 57, 59, 60, 62, 68, 72, 80 e 91. O posicionamento de cada distrator nos grupos foi feito sempre de acordo com o grupo de maior proficiência em que foi observada uma probabilidade de marcação acima de 0,5. Utilizando essa metodologia, se chegou ao resultado apresentado na Figura 13.

Figura 13 - Posicionamento das questões/opções

	Q82_B							
	Q79_B							
	Q78_B	Q90_B						
	Q77_C	Q89_B						
	Q76_D	Q86_A						
	Q75_C	Q85_B						
	Q69_B	Q83_A						
	Q67_A	Q65_A						
	Q53_C	Q64_B						
	Q50_B	Q63_B						
	Q45_A	Q61_A						
	Q44_B	Q58_B						
	Q42_B	Q56_C						
	Q41_C	Q55_D						
	Q40_B	Q52_B						
Q87_B	Q39_C	Q49_D						
Q84_B	Q37_D	Q48_B						
Q81_A	Q32_C	Q43_B						
Q74_D	Q31_A	Q35_B						
Q73_A	Q29_B	Q34_A						
Q71_D	Q26_C	Q33_B						
Q66_B	Q20_A	Q22_A						
Q54_A	Q19_B	Q15_C						
Q51_C	Q17_B	Q12_D						
Q27_C	Q16_B	Q11_B	Q46_A					
Q25_D	Q14_C	Q10_D	Q38_A					
Q23_C	Q13_B	Q9_C	Q25_A	Q70_C				
Q18_C	Q8_B	Q3_C	Q7_A	Q21_B				
Q1_A	Q5_A	Q2_D	Q4_C	Q6_A	Q71_A			
Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5	Grupo 6	Grupo 7	Grupo 8	Grupo 9
(0,12]	(12,19]	(19,23]	(23,31]	(31,35]	(35,44]	(44,52]	(52,68]	(68,100]

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

A interpretação dessa figura revela que, para os seis primeiros grupos de proficiência, é possível identificar pelo menos um erro característico de cada nível de desempenho, o que pode contribuir para a construção de um perfil de erros comuns para esses intervalos de valores de proficiência. O comportamento do posicionamento das opções, assim como na construção de uma escala de proficiência que tem como base os acertos dos estudantes, se assemelha a uma curva de

Gauss[13], porém, deslocada à esquerda, o que se explica pelo fato de que o foco de atenção da análise aqui apresentada é justamente a população com valores de proficiência baixos, ou seja, aqueles que mais cometem erros.

Consequentemente, se observa no extremo oposto da escala (valores de proficiência maiores que 44), que não há opções erradas a serem posicionadas, o que parece indicar que essa população já não comete mais erros de forma consistente, que permita a caracterização de um perfil. Isso não significa que esses indivíduos não cometam erro no teste, mas sim que esses erros não são majoritários e, portanto, não podem ser estatisticamente considerados para gerar evidências sobre o seu desempenho. Esse resultado também é coerente com os objetivos da presente pesquisa, que não se detém na análise do desempenho de estudantes com valores de proficiência elevados.

Partindo para a interpretação dessa escala, foi realizada a descrição de cada opção que seguiu o critério de pelo menos 0,5 de probabilidade de marcação entre os estudantes de determinado nível. Essa descrição foi feita para o nível mais alto em que esse percentual foi observado. O Quadro 11, a seguir, mostra essas informações.

[13]A curva de distribuição normal, também chamada de curva de Gauss, consiste em um gráfico de densidade utilizado como modelo para representar o comportamento de fenômenos aleatórios com base em dois parâmetros: média e desvio-padrão (SALKIND, 2007). A distribuição esperada dos itens e dos indivíduos prevê cerca de 68% concentrados a um desvio padrão da média, tanto para cima quanto para baixo. Dentro de dois desvios padrões para cima e para baixo, estariam aproximadamente 95% dos indivíduos e dos itens, e no espaço de 3 desvios padrões estaria quase a totalidade da população alvo e dos itens aplicados.

Quadro 11 - Descrição das opções posicionadas na escala (continua)

Questão	Nível	Descrição do erro
Q1	(0-12]	Faz interpretação equivocada de tirinha considerando elementos visuais de forma isolada, demonstrando não ser capaz de reconhecer efeito de sentido de expressão no rosto de personagem.
Q2	(19-23]	Seleciona informação que é a primeira a ser apresentada no texto, dentre as opções.
Q3	(19-23]	Associa significado de palavra a outra que aparece com maior frequência no texto.
Q4	(23-31]	Identifica palavra que faz parte do cotidiano infantil, e não uma que caracteriza a linguagem de jovens.
Q5	(12-19]	Faz interpretação equivocada de tirinha considerando elementos visuais de forma isolada.
Q6	(31-35]	Faz interpretação literal e isolada do primeiro quadrinho de tirinha, desconsiderando o restante da história.
Q7	(23-31]	Faz interpretação equivocada de expressão em conto.
Q8	(12-19]	Associa informações do texto com base em interpretação equivocada
Q9	(19-23]	Problema na elaboração do item
Q10	(19-23]	Estabelece relação de causa/consequência com base em conhecimento de mundo, e não com base no texto
Q11	(19-23]	Seleciona informação que é a primeira a ser apresentada no texto, dentre as opções.
Q12	(19-23]	Responde com base em conhecimento de mundo, e não no texto.
Q13	(12-19]	Problema na elaboração do item
Q14	(12-19]	Confunde sequência cronológica de informações do texto e faz interpretação equivocada.
Q15	(19-23]	Problema na elaboração do item
Q16	(12-19]	Entende a expressão "saiu errado" como uma gíria.
Q17	(12-19]	Faz interpretação equivocada do texto.
Q18	(0-12]	Problema na elaboração do item
Q19	(12-19]	Problema na elaboração do item
Q20	(12-19]	Confunde o espaço de história ao fazer inferência equivocada a partir de trecho do texto.
Q21	(31-35]	Associa fala em texto a personagem e não ao narrador, demonstrando não ser capaz de reconhecer fala de narrador cujo nome não é mencionado no texto.
Q22	(19-23]	Não reconhece narrador onisciente em textos.
Q23	(0-12]	Identifica informação explícita em tirinha, demonstrando não ser capaz de interpretar efeito de humor
Q25	(0-12]	Não reconhece onomatopeia em tirinha, associando-a equivocadamente a uma fala em língua estrangeira

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Quadro 11 - Descrição das opções posicionadas na escala (continua)

Questão	Nível	Descrição do erro
Q25	(23-31]	Localiza informação explícita em tirinha, demonstrando não ser capaz de inferir sentido do uso de onomatopeia
Q26	(12-19]	Faz interpretação de charge com base em elementos visuais isolados, demonstrando não ser capaz de reconhecer efeito de sentido de expressão no rosto de personagem.
Q27	(0-12]	Estabelece significado somente a partir de imagem, não fazendo relação com texto verbal.
Q29	(12-19]	Seleciona informação que é a primeira a ser apresentada no corpo do texto, dentre as opções.
Q31	(12-19]	Problema na elaboração do item
Q32	(12-19]	Responde com base em conhecimento de mundo, e não no texto.
Q33	(19-23]	Problema na elaboração do item
Q34	(19-23]	Faz interpretação equivocada de informação do início do texto.
Q35	(19-23]	Não é capaz de reconhecer trecho que caracteriza o humor em tirinha, selecionando outro cujo conteúdo irá desencadear o efeito de humor.
Q37	(12-19]	Problema na elaboração do item
Q38	(23-31]	Seleciona informação que é a primeira a ser apresentada no texto, dentre as opções.
Q39	(12-19]	Não compreende efeito de humor, registrando apenas informação visual explícita de tirinha
Q40	(12-19]	Não compreende tirinha, associando imagem isolada ao seu conhecimento de mundo.
Q41	(12-19]	Problema na elaboração do item
Q42	(12-19]	Não compreende efeito de humor, associando o humor da tirinha à expressão de alegria da personagem.
Q43	(19-23]	Confunde narrador de história com personagem que realiza as ações narradas
Q44	(12-19]	Extrapola tarefa solicitada no comando, fazendo interpretação de trecho do texto com base em conhecimento de mundo.
Q45	(12-19]	Estabelece relação entre imagem e texto verbal em tirinha, porém, considerando quadrinho diferente do solicitado no comando.
Q46	(23-31]	Seleciona informação que é a primeira a ser apresentada no texto, dentre as opções.
Q48	(19-23]	Se detém em caracterização de personagem apresentada no início do texto.
Q49	(19-23]	Faz associação equivocada de partes de um texto, demonstrando dificuldades na compreensão de recursos anafóricos
Q50	(12-19]	Associa equivocadamente voz do narrador a nome citado no texto.
Q51	(0-12]	Confere sentido a expressão a partir de interpretação global de texto escrito, e não da expressão especificamente
Q52	(19-23]	Não é capaz de reconhecer trecho que caracteriza o humor em tirinha, selecionando outro cujo conteúdo irá desencadear o efeito de humor.

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Quadro 11 - Descrição das opções posicionadas na escala (continua)

Questão	Nível	Descrição do erro
Q53	(12-19]	Falha na compreensão de aspectos visuais de tirinha, fazendo uma interpretação equivocada
Q54	(0-12]	Não é capaz de reconhecer expressão de emoções mais complexas em tirinha, associando expressão de contrariedade a tristeza.
Q55	(19-23]	Não é capaz de reconhecer trecho que caracteriza o humor em tirinha, selecionando outro cujo conteúdo irá desencadear o efeito de humor.
Q56	(19-23]	Associa equivocadamente expressão a palavra próxima no texto.
Q58	(19-23]	Confere sentido a palavra a partir de interpretação global de texto escrito, e não da palavra especificamente
Q61	(19-23]	Faz associação com base em conhecimento de mundo, e não no texto.
Q63	(19-23]	Problema na elaboração do item
Q64	(19-23]	Faz associação equivocada de partes de um texto, demonstrando dificuldades na sua compreensão
Q65	(19-23]	Escolhe opção que tem palavra igual ao que é solicitado no comando.
Q66	(0-12]	Estabelece relação de causa/consequência a partir de inferência que extrapola o conteúdo de texto escrito.
Q67	(12-19]	Desconsidera texto verbal na interpretação de imagem em quadrinho.
Q69	(12-19]	Interpreta frase de maneira isolada do texto, demonstrando dificuldades em compreensão leitora
Q70	(31-35]	Não é capaz de interpretar tirinha, atendo-se a informações explícitas.
Q71	(0-12]	Faz interpretação equivocada do texto, extrapolando seu conteúdo.
Q71	(35-44]	Faz associação equivocada do texto a partir de informação explícita do texto.
Q73	(0-12]	Não reconhece pessoas que dialogam como personagens de texto.
Q74	(0-12]	Extrapola conteúdo do texto, imaginando final feliz para história.
Q75	(12-19]	Não é capaz de interpretar texto, escolhendo opção que tem palavra igual no texto.
Q76	(12-19]	Faz leitura de elementos visuais isolados de tirinha, interpretando-a equivocadamente.
Q77	(12-19]	Faz interpretação literal de anedota com base em conhecimento de mundo, demonstrando não ser capaz de reconhecer efeito de humor.
Q78	(12-19]	Estabelece relação de causa/consequência em poema a partir de seu conhecimento de mundo, e não do texto.
Q79	(12-19]	Estabelece relação de causa/consequência em tirinha exclusivamente a partir das imagens, ignorando os balões de fala
Q81	(0-12]	Confunde fala de narrador com descrição de personagem, inferindo informação incondizente com texto escrito.
Q82	(12-19]	Entende frase como sendo fala de personagem que é mencionado explicitamente em seguida no texto.
Q83	(19-23]	Faz interpretação literal de expressão popular.

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

Quadro 11 - Descrição das opções posicionadas na escala (conclusão)

Questão	Nível	Descrição do erro
Q84	(0-12]	Entende frase em 1ª pessoa como registro de linguagem informal.
Q85	(19-23]	Associa equivocadamente significado de palavra a expressão mencionada logo antes no texto.
Q86	(19-23]	Seleciona informação que é a primeira a ser apresentada no texto, dentre as opções.
Q87	(0-12]	Faz interpretação literal de uso de gíria, não reconhecendo efeito de humor em tirinha.
Q89	(19-23]	Faz interpretação equivocada do texto, conferindo à palavra sentido oposto.
Q90	(19-23]	Faz interpretação equivocada de tirinha, atendo-se a palavras isoladas.

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

No processo de interpretação pedagógica dos itens, se verificou que 10 questões apresentavam problemas de elaboração, motivo pelo qual elas não foram descritas. Tais problemas estão descritos no Quadro 12, a seguir.

Quadro 12 - Problemas de elaboração encontrados

Questão	Nível	Problema de elaboração
Q9	19-23	Questão apresenta opção que se destaca entre as demais, induzindo a resposta do estudante
Q13	12-19	Questão apresenta palavra provavelmente desconhecida pelos estudantes, induzindo a sua resposta
Q15	19-23	Questão apresenta opção que se destaca entre as demais por apresentar palavra retirada do texto, induzindo a resposta dos estudantes
Q18	0-12	Questão demanda o reconhecimento de siglas de estados para que o estudante chegue à resposta, extrapolando a avaliação da habilidade
Q19	12-19	Questão foi construída de modo a permitir mais de uma resposta
Q31	12-19	Questão foi construída de modo a permitir mais de uma resposta
Q33	19-23	Questão apresenta comando ambíguo
Q37	12-19	Questão apresenta opção que se destaca entre as demais, induzindo a resposta do estudante
Q41	12-19	Questão apresenta comando ambíguo e erro ortográfico no texto que o deixa também ambíguo, dificultando a interpretação do estudante
Q63	19-23	Questão apresenta pegadinha, uma vez que possui opção com palavra que induz estudante ao erro

Fonte: Elaborado pela autora (2022).
*Dados extraídos da Prova São Paulo 2018.

É importante também destacar a interpretação das questões 25 e 71, que apresentaram resultados que se aproximam do comportamento esperado quanto à marcação de opções erradas por estudantes que ainda não desenvolveram a habilidade que está sendo avaliada.

Como mencionado anteriormente, nessas questões verificou-se que duas opções tiveram probabilidade de marcação acima de 0,5 para níveis diferentes da escala. Na primeira, estudantes do grupo 1 apresentaram maior probabilidade de cometer o erro atribuído à opção D e estudantes dos grupos 3 e 4, o erro atribuído à opção A. Já na questão 71, estudantes do grupo 1 tenderam a cometer o erro relacionado à opção D, enquanto a opção A foi escolha da maioria dos estudantes dos grupos 2 a 6. A descrição dos erros de ambas as questões é retomada no Quadro 13, a seguir.

Quadro 13 - Descrição dos erros verificados nas questões 25 e 71

Questão	Nível	Proposta de descrição do erro
Q25	(0-12]	Não reconhece onomatopeia em tirinha, associando-a equivocadamente a uma fala em língua estrangeira
Q25	(23-31]	Localiza informação explícita em tirinha, demonstrando não ser capaz de inferir sentido do uso de onomatopeia
Q71	(0-12]	Faz interpretação equivocada do texto, extrapolando seu conteúdo.
Q71	(35-44]	Faz associação equivocada do texto a partir de informação explícita do texto.

O conteúdo dos erros descritos permite inferir, para essas questões, que estudantes do grupo 1 cometem erros mais simples do que os associados aos grupos subsequentes, havendo uma certa progressão entre eles. No caso da questão 25, enquanto o grupo 1 não é capaz sequer de reconhecer uma onomatopeia, confundindo-a com língua estrangeira, os grupos que o sucedem, apesar de não conseguirem inferir o sentido do uso da onomatopeia, são capazes, pelo menos, de reconhecê-la. Já na questão 71, enquanto os estudantes do grupo 1 não conseguem interpretar o texto, aqueles com proficiências ligeiramente maiores já são capazes de interpretar o texto, fazendo associações, apesar de equivocadas, a partir de informações explícitas.

Tal comportamento reforça o que vem sendo defendido neste estudo: de que existe um *continuum* no processo de construção habilidades pelos estudantes, o que pode e deve ser considerado durante a elaboração de itens de avaliação em larga escala.

As propostas de interpretação dos erros, em conjunto, permitem construir uma escala interpretada dos erros cometidos pelos estudantes no teste de língua portuguesa da Prova São Paulo 2018, documento que compõe o produto técnico da presente dissertação e será apresentado na próxima sessão.

Considerações finais

O presente estudo investigou metodologias de análise que permitam considerar, além dos acertos, os erros dos estudantes ao realizarem testes de avaliações em larga escala. Com isso, espera-se somar à análise que atualmente é feita dos resultados dessas avaliações, informações a respeito dos erros que os estudantes tendem a cometer durante o teste, buscando ajudar escolas e

redes de ensino a melhor compreender o desempenho dos seus estudantes e a melhor intervir para que os resultados sejam cada vez mais positivos.

Nessa investigação, optou-se por adotar os modelos MRN e MRG da TRI que, segundo pesquisas realizadas, seriam aqueles que poderiam oferecer o tipo de análise ao qual esta pesquisa se propôs. Utilizando essa metodologia, foi possível calcular os parâmetros *a* e *b* para os itens do *corpus* selecionado e, assim, construir uma escala, em que as opções dos itens foram posicionadas de acordo com a habilidade (ou valor de proficiência) dos estudantes.

É importante frisar, porém, que essa é uma análise exploratória, que precisa ser aprofundada em termos de rigor estatístico para verificar se de fato os modelos adotados são os mais adequados. Ademais, a investigação do erro no universo das avaliações em larga escala é ainda inicial, motivo pelo qual carece de referencial teórico e de mais pesquisas para consolidar os seus resultados.

Independente disso, os resultados alcançados nesta pesquisa mostram que a análise dos erros dos estudantes em avaliações em larga escala é possível e, tendo em vista os resultados que oferece, merece ser considerada como mais uma possibilidade de interpretação dos seus dados resultantes, ajudando a "identificar erros comuns que os professores devem conhecer e tentar solucionar. Esses materiais podem servir de base para os atores discutirem os resultados e desenvolverem estratégias para abordar áreas de baixo desempenho" (OCDE, 2021, p. 70). Assim, teremos condições de extrair ainda mais informações dos dados que são coletados, contribuindo com ferramentas que ajudem professores e gestores escolares a melhor compreender o desempenho dos seus estudantes nessas avaliações.

Referências

ARAUJO, E. A. C. de; ANDRADE, D. F. de; BORTOLOTTI, S. L. V. Teoria da Resposta ao Item. **Rev Esc Enferm USP**, São Paulo, v. 43, n. esp., p. 1000-1008, 2009. Disponível em: <https://www.scielo.br/j/reeusp/a/VS9FdSVm6CsSxQYkJ5nr8tD/?lang=pt>. Acesso em 02 jul. 2022.

BRASIL. **Parâmetros curriculares nacionais**: introdução aos parâmetros curriculares nacionais. Secretaria de Educação Fundamental. Brasília: MEC/SEF, 1997. Disponível em: <http://portal.mec.gov.br/component/content/article?id=12640:parametros-curriculares-nacionais-1o-a-4o-series>. Acesso em 02 jul. 2022.

CHAPPAZ, R. de O.; ALAVARSE, O. M. **Avaliação externa na Rede Municipal de Ensino de São Paulo**: os desafios da participação docente. Cadernos Cenpec, São Paulo, v.7, n.2, p.88-111, jul./dez. 2017. Disponível em: <https://cadernos.cenpec.org.br/cadernos/index.php/cadernos/article/view/402>. Acesso em 02 jul. 2022.

FRANÇA, M. J. F. **Avaliação em larga escala**: um estudo sobre erros dos alunos no trabalho com números e suas operações. 2008. 115 f. Dissertação (Mestrado em Educação). Universidade Federal do Pernambuco, Recife, 2008. Disponível em: <https://bdtd.ibict.br/vufind/Record/UFPE_b8c6642eb115af7898c9ba9750387a70>. Acesso em 02 jul. 2022.

GONÇALVES, A. **Análise das estratégias e erros dos alunos do 9º ano em questões de álgebra baseadas no Saresp de 2008 a 2011**. 2014. 178 f. Dissertação (Mestrado em Educação

Matemática). Pontifícia Universidade Católica de São Paulo, São Paulo, 2014. Disponível em: <https://repositorio.pucsp.br/jspui/handle/handle/10988>. Acesso em 02 jul. 2022.

HAIR JR., J. F.; TATHAM, R. L.; ANDERSON, R. E.; BLACK, W. Análise multivariada de dados. 6. ed. Porto Alegre: Bookman, 2009.

LUCKESI, C. C. **O que é mesmo o ato de avaliar a aprendizagem**. Pátio, Porto Alegre, ano 3, n. 12, fev./abr. 2000. Disponível em: <https://www.nescon.medicina.ufmg.br/biblioteca/imagem/2511.pdf>. Acesso em 02 jul. 2022.

NOGARO, A.; GRANELLA, E. O erro no processo de ensino e aprendizagem. **Revista de Ciências Humanas**, Frederico Westphalen, v. 5, n. 5, p. 31-56, 2004. Disponível em: <http://revistas.fw.uri.br/index.php/revistadech/article/view/244/445>. Acesso em 02 jul. 2022.

OCDE. Reforma da Avaliação Nacional: principais considerações para o Brasil. Perspectivas da OCDE sobre políticas educacionais. Tradução de Marília Aranha. Disponível em: <https://fundacaolemann.org.br/storage/materials/vDuLwKGNyktGQRiSnKxemSWs7fm3wZlHljYgO8jV.pdf>. Acesso em: 27 jun. 2022.

PINHEIRO, I. R.; COSTA, F. R.; CRUZ, R. M. Modelo nominal da Teoria de Resposta ao Item: uma alternativa. **Avaliação psicológica**, Porto Alegre, v. 9, n. 3, p. 437-447, dez. 2010. Disponível em <http://pepsic.bvsalud.org/scielo.php?script=sci_abstract&pid=S1677-04712010000300010&lng=pt&nrm=iso>. Acesso em: 02 jul. 2022.

ROSSO, A. J.; BERTI, N. M. O erro e o ensino-aprendizagem de matemática na perspectiva do desenvolvimento da autonomia do aluno. **Bolema**, Rio Claro, v. 23, n. 37, p. 1005-1035, dez. 2010. Disponível em:

<https://www.periodicos.rc.biblioteca.unesp.br/index.php/bolema/article/view/4313>. Acesso em 02 jul. 2022.

SALKIND, N. J. (ed.) **Encyclopedia of measurement and statistics**. v. 1. Thousand Oaks: SAGE Publications, 2007. Disponível em: <https://dx.doi.org/10.4135/9781412952644>. Acesso em 21 ago. 2021.

SAMEJIMA, F. (1969). Estimation of latent ability using a response pattern of graded scores (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Disponível em: <https://www.psychometricsociety.org/sites/main/files/file-attachments/mn17.pdf>. Acesso em: 27 jun. 2022.

SANTOS, C. I. dos; OLIVEIRA, P. C. Avaliação externa em matemática: análise de teses e dissertações que abordam conteúdos matemáticos. **Rev. Bras. de Iniciação Científica (RBIC)**, Itapetininga, v. 7, n. 3, p. 36-55, abr./jun. 2020. Disponível em: <https://periodicos.itp.ifsp.edu.br/index.php/IC/article/view/1478>. Acesso em 02 jul. 2022.

SÃO PAULO. **Lei nº 14.063, de 14 de outubro de 2005**. Institui o Sistema de Avaliação de Aproveitamento Escolar dos Alunos da Rede Municipal de Ensino de São Paulo, sob a responsabilidade da Secretaria Municipal de Educação. Disponível em: <http://legislacao.prefeitura.sp.gov.br/leis/lei-14063-de-14-de-outubro-de-2005>. Acesso em 03 out. 2021.

SÃO PAULO. **Portaria Secretaria Municipal de Educação - SME nº 2.639 de 10 de março de 2017**. Orienta a aplicação das Avaliações Externas integrantes do Sistema de Avaliação Escolar dos Alunos da RME e dá outras providências. Disponível em: <

http://legislacao.prefeitura.sp.gov.br/leis/portaria-secretaria-municipal-de-educacao-sme-2639-de-10-de-marco-de-2017>. Acesso em 27 ago. 2022.

SCHUBRING, G. Desenvolvimento histórico do conceito e do processo de aprendizagem, a partir de recentes concepções matemático-didáticas (erro, obstáculos, transposição). Tradução de Pedro Goergen. **Zetetiké**, Campinas, v. 6, n. 10, p. 9-34, jul./dez. 1998. Disponível em: <https://doi.org/10.20396/zet.v6i10.8646782>. Acesso em 02 jul. 2022.

SILVA, E. M. D. da. A virtude do erro: uma visão construtivista da avaliação. **Estudos em Avaliação Educacional**, São Paulo, v. 19, n. 39, p. 91-114, jan./abr. 2008. Disponível em: <https://oasisbr.ibict.br/vufind/Record/FCC-2_976e5856ee4845a183d80433291e8adf>. Acesso em 02 jul. 2022.

SOUSA, M. de; FERRAROTTO, L. **Avaliação externa na Rede Municipal de São Paulo**: para qual direção conduz? Rev. Teoria e Prática da Educação, v. 19, n. 3, p. 79-89, set./dez. 2016. Disponível em: <https://periodicos.uem.br/ojs/index.php/TeorPratEduc/article/view/36622>. Acesso em 02 jul. 2022.

SPINILLO, A. G. et al. O erro no processo de ensino-aprendizagem da matemática: errar é preciso? **Boletim Gepem (Online)**, Rio de Janeiro, n. 64, p. 1-15, jan./jun. 2014. Disponível em: <https://periodicos.ufrrj.br/index.php/gepem/article/view/13>. Acesso em 02 jul. 2022.

PRODUTO TÉCNICO

Como parte integrante da dissertação no Programa de Pós-graduação em Educação – Modalidade Profissional (PPGE-MP), é requerida a apresentação de um produto técnico, que tem como objetivo fazer a ponte entre a pesquisa científica e o contexto do trabalho. Dessa maneira, espera-se promover o aprimoramento profissional com base em estudos teóricos que gerem impacto nas realidades vivenciadas.

No presente estudo, o produto técnico proposto é a escala interpretada dos erros cometidos pelos estudantes em uma avaliação em larga escala, apresentada no Quadro 14. Espera-se, com esse documento, oferecer à comunidade escolar e à sociedade em geral uma devolutiva de caráter pedagógico a respeito do desempenho dos estudantes nessa avaliação, além de gerar evidências de intervenção na prática pedagógica.

Muito se argumenta sobre a ineficiência das avaliações em larga escala em trazer informações que de fato contribuam para as práticas pedagógicas do professor (PERRY, 2009; SOUSA; OLIVEIRA, 2010; KISTEMANN JR; GOUVÊA, 2019; CALDERÓN; BORGES, 2020). Dados são coletados, testes são respondidos, proficiências são calculadas, mas não se chega a uma interpretação adequada das informações, de modo que o professor, ao se debruçar sobre elas, não consegue compreender o desempenho dos seus alunos e os aspectos nos quais ele deve empreender maiores esforços (FONTANIVE; ELLIOT; KLEIN, 2007; SOUSA; OLIVEIRA, 2010).

Fica evidente, aí, portanto, um vácuo entre a realização de uma avaliação em larga escala e a aplicabilidade dos seus resultados na rotina escolar, uma vez que a interpretação que deveria ser feita de seus dados não parece estar ao alcance de gestores e professores ou não é detalhada o suficiente para que eles possam aplicá-la em sua práxis.

É visando preencher essa lacuna que o produto técnico da presente dissertação se insere, ao propor a interpretação pedagógica dos resultados de uma avaliação em larga escala. Essa interpretação, porém, ao invés de levar em consideração os acertos dos estudantes em um teste – metodologia demonstrada no estudo I da presente dissertação –, tem como referência os seus erros nesse mesmo teste. Assim, espera-se contribuir adicionando detalhamento à divulgação dos

resultados e colaborando, em última instância, para uma melhor compreensão dos resultados de uma avaliação em larga escala.

Essa interpretação dos erros parte de uma análise estatística dos resultados considerando as opções marcadas pelos estudantes que não correspondem ao gabarito do item, abordada no estudo II desta dissertação. Por essa razão, focaliza os estudantes que mais cometeram erros e que, portanto, tiveram um desempenho fraco ou ruim no teste.

Destaca-se que a escolha pela análise dos erros e por esse perfil de estudante foi feita partindo-se do entendimento de que são justamente esses os dois aspectos que mais devem ser enfatizados ao reportar resultados de uma avaliação em larga escala para professores: os estudantes que tiveram desempenho ruim e os erros que foram cometidos. Parte-se do pressuposto de que, tendo conhecimento dessas duas variáveis, o professor terá mais condições e ferramentas para compreender os resultados da avaliação e reorientar suas práticas para alcançar melhores resultados.

Quadro 14 - Escala interpretada dos erros dos estudantes do 5º ano em língua portuguesa, na Prova São Paulo 2018 (continua)

Nível 0-12	Um estudante com proficiência menor que 12 no teste tende a cometer os seguintes erros: • confere sentido a expressão a partir de interpretação global de texto escrito, e não da expressão especificamente; • confunde fala de narrador com descrição de personagem, inferindo informação incondizente com texto escrito; • entende frase em 1ª pessoa como registro de linguagem informal; • estabelece relação de causa/consequência a partir de inferência que extrapola o conteúdo de texto escrito; • estabelece significado somente a partir de imagem, não fazendo relação com texto verbal; • extrapola conteúdo do texto, imaginando final feliz para história; • faz interpretação equivocada de tirinha considerando elementos visuais de forma isolada, demonstrando não ser capaz de reconhecer efeito de sentido de expressão no rosto de personagem; • faz interpretação equivocada do texto, extrapolando seu conteúdo; • faz interpretação literal de uso de gíria, não reconhecendo efeito de humor em tirinha; • identifica informação explícita em tirinha, demonstrando não ser capaz de interpretar efeito de humor; • não é capaz de reconhecer expressão de emoções mais complexas em tirinha, associando expressão de contrariedade a tristeza; • não reconhece onomatopeia em tirinha, associando-a equivocadamente a uma fala em língua estrangeira; e • não reconhece pessoas que dialogam como personagens de texto.
Nível 12-19	Um estudante com proficiência maior ou igual a 12 e menor que 19 no teste tende a cometer os seguintes erros: • associa equivocadamente voz do narrador a nome citado no texto; • associa informações do texto com base em interpretação equivocada; • confunde o espaço de história ao fazer inferência equivocada a partir de trecho do texto; • confunde sequência cronológica de informações do texto e faz interpretação equivocada; • desconsidera texto verbal na interpretação de imagem em quadrinho; • entende a expressão “saiu errado” como uma gíria; • entende frase como sendo fala de personagem que é mencionado explicitamente em seguida no texto; • estabelece relação de causa/consequência em poema a partir de seu conhecimento de mundo, e não do texto; • estabelece relação de causa/consequência em tirinha exclusivamente a partir das imagens, ignorando os balões de fala;

Quadro 14 - Escala interpretada dos erros dos estudantes do 5º ano em língua portuguesa, na Prova São Paulo 2018 (continua)

Nível 12-19	• estabelece relação entre imagem e texto verbal em tirinha, porém, considerando quadrinho diferente do solicitado no comando; • extrapola tarefa solicitada no comando, fazendo interpretação de trecho do texto com base em conhecimento de mundo. • falha na compreensão de aspectos visuais de tirinha, fazendo uma interpretação equivocada; • faz interpretação de charge com base em elementos visuais isolados, demonstrando não ser capaz de reconhecer efeito de sentido de expressão no rosto de personagem; • faz interpretação equivocada de tirinha considerando elementos visuais de forma isolada; • faz interpretação equivocada do texto; • faz interpretação literal de anedota com base em conhecimento de mundo, demonstrando não ser capaz de reconhecer efeito de humor; • faz leitura de elementos visuais isolados de tirinha, interpretando-a equivocadamente. • interpreta frase de maneira isolada do texto, demonstrando dificuldades em compreensão leitora; • não compreende efeito de humor, associando o humor da tirinha à expressão de alegria da personagem; • não compreende efeito de humor, registrando apenas informação visual explícita de tirinha; • não compreende tirinha, associando imagem isolada ao seu conhecimento de mundo. • não é capaz de interpretar texto, escolhendo opção que tem palavra igual no texto. • responde com base em conhecimento de mundo, e não no texto; e • seleciona informação que é a primeira a ser apresentada no corpo do texto, dentre as opções.
Nível 19-23	Um estudante com proficiência maior ou igual a 19 e menor que 23 no teste tende a cometer os seguintes erros: • associa equivocadamente expressão a palavra próxima no texto; • associa equivocadamente significado de palavra a expressão mencionada logo antes no texto; • associa significado de palavra a outra que aparece com maior frequência no texto; • confere sentido a palavra a partir de interpretação global de texto escrito, e não da palavra especificamente; • confunde narrador de história com personagem que realiza as ações narradas; • estabelece relação de causa/consequência com base em conhecimento de mundo, e não com base no texto; • faz associação com base em conhecimento de mundo, e não no texto;

Quadro 14 - Escala interpretada dos erros dos estudantes do 5º ano em língua portuguesa, na Prova São Paulo 2018 (continua)

Nível 19-23	• faz associação equivocada de partes de um texto, demonstrando dificuldades na compreensão de recursos anafóricos; • faz associação equivocada de partes de um texto, demonstrando dificuldades na sua compreensão; • faz interpretação equivocada de informação do início do texto; • faz interpretação equivocada de tirinha, atendo-se a palavras isoladas; • faz interpretação equivocada do texto, conferindo à palavra sentido oposto; • faz interpretação literal de expressão popular; • não é capaz de reconhecer trecho que caracteriza o humor em tirinha, selecionando outro cujo conteúdo irá desencadear o efeito de humor; • não reconhece narrador onisciente em textos; • responde com base em conhecimento de mundo, e não no texto; • se detém em caracterização de personagem apresentada no início do texto; e • seleciona informação que é a primeira a ser apresentada no texto, dentre as opções.
Nível 23-31	Um estudante com proficiência maior ou igual a 23 e menor que 31 no teste tende a cometer os seguintes erros: • faz interpretação equivocada de expressão em conto. • identifica palavra que faz parte do cotidiano infantil, e não uma que caracteriza a linguagem de jovens. • localiza informação explícita em tirinha, demonstrando não ser capaz de inferir sentido do uso de onomatopeia • seleciona informação que é a primeira a ser apresentada no texto, dentre as opções.
Nível 31-35	Um estudante com proficiência maior ou igual a 31 e menor que 35 no teste tende a cometer os seguintes erros: • associa fala em texto a personagem e não ao narrador, demonstrando não ser capaz de reconhecer fala de narrador cujo nome não é mencionado no texto. • faz interpretação literal e isolada do primeiro quadrinho de tirinha, desconsiderando o restante da história. • não é capaz de interpretar tirinha, atendo-se a informações explícitas.
Nível 35-44	Um estudante com proficiência maior ou igual a 35 e menor que 44 no teste tende a cometer o seguinte erro: • faz associação equivocada a partir de informação explícita do texto.

Considerações finais

A proposta de elaboração de uma escala interpretada que leve em consideração o erro do estudante na aplicação de um teste de língua portuguesa busca atender às exigências de desenvolvimento do produto técnico do Mestrado Profissional da Faculdade de Educação da Universidade de Brasília.

Para o desenvolvimento desse produto, foram utilizados os resultados obtidos a partir da análise estatística e pedagógica dos itens de língua portuguesa aplicados aos estudantes do 5º ano do ensino fundamental na edição de 2018 da Prova São Paulo. Com a análise estatística, foi possível definir uma escala, em que as opções erradas com maior probabilidade de marcação foram posicionadas de acordo com os grupos de valores de proficiência estabelecidos. Já a análise pedagógica permitiu a interpretação desses erros a partir da descrição de cada opção contida na escala, buscando construir um perfil de erros cometidos na resolução da referida prova de acordo com a proficiência dos estudantes que responderam ao teste.

Destaca-se que a proposta dessa escala consiste em um estudo preliminar sobre o uso de ferramentas estatísticas para a análise do erro em testes cognitivos. Sendo assim, não pretende esgotar a temática e os procedimentos abordados, tampouco desenvolver um método definitivo para a condução de análises no mesmo modelo. Se almeja unicamente oferecer possibilidades de ampliação do uso da metodologia de desenvolvimento de escalas de proficiência que contribuam para um melhor aproveitamento dos dados coletados em uma avaliação em larga escala.

Espera-se que a presente proposta de escala interpretada do erro possa oferecer evidências pedagógicas que auxiliem professores a compreender o desempenho de estudantes nessas avaliações, e assim possam atuar para melhorar a qualidade do ensino.

Referências

PERRY, F. A. **Escalas de proficiência**: diferentes abordagens de interpretação na avaliação educacional em larga escala. Juiz de Fora, 2009. 119 f. Dissertação (Mestrado em Educação) - Programa de Pós-Graduação em Educação, Universidade Federal de Juiz de Fora, 2009. Disponível em: <https://repositorio.ufjf.br/jspui/handle/ufjf/3835>. Acesso em: 21 ago. 2021.

SOUSA, S. Z. L.; OLIVEIRA, R. P. de. Sistemas estaduais de avaliação: uso dos resultados, implicações e tendências. **Cadernos de Pesquisa**, São Paulo, v. 40, n. 141, p. 793-822, set./dez. 2010. Disponível em: <https://www.scielo.br/j/cp/a/HfYnBHFv4x63bWY6nkfJt7H/abstract/?lang=pt>. Acesso em: 21 ago. 2021.

KISTEMANN JR, M. A.; GOUVÊA, C. de L. Uma investigação com professores de matemática e sua leitura dos resultados das avaliações em larga escala (Proeb). **Revista Pesquisa e Debate em Educação**, Juiz de Fora, v. 9, n. 1, p. 606-624, 2019. Disponível em: <https://doi.org/10.34019/2237-9444.2019.v9.31132>. Acesso em: 21 ago. 2021.

CALDERÓN, A. I.; BORGES, R. M. Avaliação em larga escala na educação básica: usos e tensões teórico-epistemológicas. **Meta: Avaliação**, Rio de Janeiro, v. 12, n. 34, p. 28-58, jan./mar. 2020. Disponível em: <http://dx.doi.org/10.22347/2175-2753v12i34.2281>. Acesso em: 21 ago. 2021.

FONTANIVE, N. S.; ELLIOT, L. G.; KLEIN, R. Os desafios da apresentação dos resultados da avaliação de sistemas escolares a diferentes públicos. **REICE - Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación**, Madri, v. 5, n. 2, p. 262-273, 2007. Disponível em: <http://hdl.handle.net/10486/660969>. Acesso em: 21 ago. 2021.

SOUSA, S. Z. L.; OLIVEIRA, R. P. de. Sistemas estaduais de avaliação: uso dos resultados, implicações e tendências. **Cadernos de Pesquisa**, São Paulo, v. 40, n. 141, p. 793-822, set./dez. 2010. Disponível em: <https://www.scielo.br/j/cp/a/HfYnBHFv4x63bWY6nkfJt7H/abstract/?lang=pt>. Acesso em: 21 ago. 2021.

OCDE. Reforma da Avaliação Nacional: principais considerações para o Brasil. Perspectivas da OCDE sobre políticas educacionais. Tradução de Marília Aranha. Disponível em: <https://fundacaolemann.org.br/storage/materials/vDuLwKGNyktGQRiSnKxemSWs7fm3wZlHljYgO8jV.pdf>. Acesso em: 17 jul. 2022.

Considerações finais da dissertação

A dissertação aqui apresentada propôs uma análise do erro como ferramenta para compreender o desempenho de estudantes em avaliações em larga escala. O desenvolvimento dessa pesquisa abrangeu a elaboração de dois estudos e um produto técnico, que compuseram a versão final do presente trabalho.

No Estudo I, foi realizada uma análise de caráter bibliográfico de documentos e relatórios técnicos publicados pelo Inep com vistas a verificar de que maneira têm sido concebidas e conduzidas as escalas de proficiência do Saeb, avaliação em larga escala de maior destaque no cenário nacional. A escolha pela investigação desse documento se deu em virtude do objetivo do presente trabalho ser o de construir uma escala nos moldes dessa citada que levasse em consideração, ao invés do acerto, como geralmente é feito, os erros cometidos pelos estudantes nos testes cognitivos que compreendem essas avaliações.

Para tanto, além de investigar os pressupostos metodológicos e teorias envolvidas na criação de uma escala de proficiência, foram analisadas as escalas publicadas desde 1995, ano em que começaram a ser construídas. Essa análise permitiu não só entender os processos que embasam a elaboração do referido documento, como também o histórico de evoluções pelas quais a escala do Saeb passou desde então. Constatou-se que, apesar de em algumas edições terem sido feitas alterações que inviabilizaram um diagnóstico temporal do desempenho dos estudantes avaliados, tais mudanças foram importantes para o aprimoramento da medida e a consolidação dos procedimentos utilizados pelo Inep para divulgar os resultados do Saeb.

A robustez alcançada ao longo desses anos foi fundamental para a criação de uma avaliação sólida que hoje é referência para diversas outras iniciativas de avaliação de caráter estadual e municipal.

O segundo estudo centrou-se no propósito de investigar metodologias que permitissem a análise estatística do erro e a construção de uma escala interpretada. Para alcançar esse objetivo, foram utilizados dados provenientes da Prova São Paulo 2018, uma iniciativa de avaliação em larga escala municipal que se espelha no Saeb para medir o desempenho dos estudantes da rede de ensino do município de São Paulo em língua portuguesa e matemática. A opção pelo uso desses

dados se deve ao fato de que, mediante solicitação enviada à Secretaria de Educação do Município de São Paulo (Anexo I), foi dado acesso a todos os dados necessários para a condução da presente pesquisa, sejam eles abertos ou sigilosos.

Nesse estudo, foram utilizados dois modelos da TRI, o MRN e o MRG, que permitem o cálculo de parâmetros para todas as opções do item, e não somente para a opção correta. Assim, foi possível analisar de um ponto de vista estatístico a adequação dessas opções e o quanto a probabilidade de marcação de uma delas estava relacionada ao nível de desempenho dos estudantes no teste, medido por meio do cálculo da proficiência desses indivíduos

A intenção, com isso, foi mapear os erros usualmente cometidos por um determinado grupo de estudantes, caracterizando os distintos perfis de desempenho na referida avaliação. A partir dessa escala, foi possível delinear, por exemplo, quais são os erros que, em geral, estudantes com determinada proficiência cometem. Em outras palavras, se até então o professor recebia a informação de quais habilidades seus estudantes demonstraram possuir ao realizarem um teste, com essa escala foi possível indicar os erros que cometeram no processo de desenvolvimento das habilidades que ainda não alcançaram.

O resultado dessa análise consistiu no produto técnico da presente dissertação, que propôs uma escala interpretada de erros de acordo com a proficiência dos estudantes que participaram da Prova São Paulo 2018. Com essa escala, foi possível apontar as deficiências que ainda caracterizam o desempenho desses estudantes e gerar evidências pedagógicas que possam ajudar professores e sistemas de ensino a melhor compreender os resultados de avaliações em larga escala e melhor aplicá-los em suas práticas. Assim, espera-se que essa escala possa contribuir, em última análise, para a melhoria da qualidade do ensino.

Os procedimentos metodológicos adotados no presente trabalho possuem caráter inovador, uma vez que não foram encontrados estudos que proponham a criação e interpretação de uma escala nos moldes aqui apresentados. Até mesmo os estudos que se centralizam nos erros dos estudantes são escassos e, quando encontrados, são da área de matemática, onde, na prática, parece haver um maior enfoque na compreensão do tipo de erro que o estudante comete. Assume-se, portanto, que o estudo apresenta limitações, principalmente na questão do embasamento teórico para as análises que aqui foram conduzidas. Ademais, reafirma-se que se trata de uma pesquisa exploratória, que não pretende esgotar tampouco consolidar os procedimentos adotados. A

dissertação pretende, na verdade, investigar possibilidades de análise dos resultados de avaliações em larga escala que ultrapassem o que já é utilizado, conhecido e tradicional, adicionando novas informações e perspectivas a respeito do desempenho dos estudantes.

Nesse propósito, o trabalho que aqui se apresenta oferece um caminho, uma possibilidade a mais para ampliar o alcance dos resultados de avaliações em larga escala e permitir uma maior apreensão dos seus impactos e significados por professores e sistemas de ensino.

Referências da dissertação

BAUER, A. Estudos sobre sistemas de avaliação educacional no Brasil: um retrato em preto e branco. **Revista @mbienteeducação**, v. 5, n. 1, p. 7-31, jan./jun. 2012.

BONAMINO, A.; FRANCO, C. Avaliação e política educacional: o processo de institucionalização do Saeb. **Cadernos de Pesquisa**, São Paulo, n. 108, p. 101-132, nov. 1999.

BONAMINO, A.; SOUSA, S. Z. Três gerações de avaliação da educação básica no Brasil: interfaces com o currículo da/na escola. **Educação e Pesquisa**, São Paulo, v. 38, n. 2, p. 373-388, abr./jun. 2012.

BRASIL. **CONSTITUIÇÃO DA REPÚBLICA FEDERATIVA DO BRASIL DE 1988** Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l9394.htm>. Acesso em: 13 mai. 2020.

BRASIL. Lei n. 13.005/2014, de 25 de junho de 2014. Aprova o Plano Nacional de Educação - PNE e dá outras providências. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l13005.htm>. Acesso em: 13 mai. 2020.

BRASIL. Lei n. 9.394/1996, de 20 de dezembro de 1996. Estabelece as diretrizes e bases da educação nacional. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l9394.htm>. Acesso em: 13 mai. 2020.

FONTANIVE, N. S.; ELLIOT, L. G.; KLEIN, R. Os desafios da apresentação dos resultados da avaliação de sistemas escolares a diferentes públicos. **REICE - Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación**, Madri, v. 5, n. 2, p. 262-273, 2007.

HORTA NETO, J. L. **As avaliações externas e seus efeitos sobre as políticas educacionais**: uma análise comparada entre a União e os Estados de Minas Gerais e São Paulo. Brasília, 2013. 358 f. Tese (Doutorado em Política Social) - Programa de Pós-Graduação em Política Social da Universidade de Brasília, 2013.

HORTA NETO, J. L. **Avaliação externa**: a utilização dos resultados do SAEB 2003 na gestão do sistema público de Ensino Fundamental no Distrito Federal. Brasília, 2006. 144 f. Dissertação (Mestrado) – Faculdade de Educação, Universidade de Brasília, 2006.

INEP. **Sistema de Avaliação da Educação Básica**: documentos de referência. Versão Preliminar. Brasília: Inep/Ministério da Educação, 2019.

KISTEMANN JR, M. A.; GOUVÊA, C. de L. Uma investigação com professores de matemática e sua leitura dos resultados das avaliações em larga escala (Proeb). **Revista Pesquisa e Debate em Educação**, Juiz de Fora, v. 9, n. 1, p. 606-624, 2019.

KLEIN, R. Como está a educação no Brasil? O que fazer? **Ensaio: aval. pol. públ. Educ.**, Rio de Janeiro, v. 14, n. 51, p. 139-172, abr./jun. 2006.

KLEIN, R. Utilização da Teoria de Resposta ao Item no Sistema Nacional de Avaliação da Educação Básica (SAEB). **Ensaio**, n. 40, v. 11, p. 283-296, jul./set. 2003.

KLEIN, R.; FONTANIVE, N. S. Avaliação em larga escala: uma proposta inovadora. **Em Aberto**, Brasília, ano 15, n. 66, p. 29-34, abr./jun. 1995.

LONGO, C. M. T. **Encontros formativos: um estudo sobre a avaliação externa e a escala de proficiência na escola de ensino integral**. 2019. 188 f. Dissertação (Mestrado em Educação: Formação de Formadores) - Programa de Estudos Pós-Graduados em Educação: Formação de Formadores, Pontifícia Universidade Católica de São Paulo, São Paulo, 2019.

MARQUES, R.; STIEG, R.; SANTOS, W. dos. Exames estandardizados: análise dos modelos e das teorias na produção acadêmica. **Meta: Avaliação**, Rio de Janeiro, v. 12, n. 34, p. 1-27, jan./mar. 2020.

OLIVEIRA, L. K. M de. **Três investigações sobre escalas de proficiência e suas interpretações**. Rio de Janeiro, 2008. 216 f. Tese (Doutorado em Educação) – Departamento de Educação, Pontifícia Universidade Católica do Rio de Janeiro, 2008.

OLIVEIRA, L. K. M. de; FRANCO, C.; SOARES, T. M. Projeto GERES/2005: novos indicadores para construção e interpretação da escala de proficiência. **REICE - Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación**, Madri, v. 5, n. 2, p. 153-182, 2007.

PERRY, F. A. **Escalas de proficiência**: diferentes abordagens de interpretação na avaliação educacional em larga escala. Juiz de Fora, 2009. 119 f. Dissertação (Mestrado em Educação) - Programa de Pós-Graduação em Educação, Universidade Federal de Juiz de Fora, 2009.

SILVA, F. S. da; LEAL, T. F. Escala de proficiência da Prova Brasil: o que informa aos professores? **Revista Leia Escola**, Campina Grande, v. 18, n. 3, p. 90-108, 2018.

SOUSA, C. P. de; FERREIRA, S. L. Avaliação em larga escala e da aprendizagem na escola: um diálogo necessário. **Psic. da Ed.**, São Paulo, n. 48, p. 13-23, 2019.

SOUSA, S. Z. L.; OLIVEIRA, R. P. de. Sistemas estaduais de avaliação: uso dos resultados, implicações e tendências. **Cadernos de Pesquisa**, São Paulo, v. 40, n. 141, p. 793-822, set./dez. 2010.

SOUZA, J. E. P. A avaliação em larga escala, o senso comum, críticas e ponderações. **Revista do Instituto de Políticas Públicas de Marília**, Marília, v. 5, n. 2, p. 139-156, jul./dez. 2019.

ZIMMARO, D. M. **Writing good multiple-choice exams**. Texas: Faculty Innovation Center, 2016.

DOWNING, S. M.; HALADYNA, T. M. Handbook of test development. New Jersey: Lawrence Earlbaum Associates, 2006.

Anexos

Carta de solicitação de acesso aos dados (continua)

Gmail — Laís Antonietto <lsantonietto@gmail.com>

Solicitação de autorização de acesso a dados - PROVA SÃO PAULO 2018

3 messages

Laís Antonietto <lsantonietto@gmail.com> Mon, Jul 20, 2020 at 5:15 PM
To: cmaroja@sme.prefeitura.sp.gov.br
Bcc: Claudia Griboski <cgriboski4@gmail.com>

Prefeitura Municipal de São Paulo

Secretaria Municipal de Educação

Coordenadoria Pedagógica – COPED

Núcleo Técnico de Avaliação – NTA

Prezado Senhor Claudio Maroja,

Meu nome é Laís Silveira Antonietto, sou colaboradora do Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos (Cebraspe) e atualmente estou vinculada como mestranda ao Programa de Pós-Graduação em Educação - Modalidade Profissional, da Universidade de Brasília (UnB), sob a orientação da Prof. Dra. Claudia M. Griboski.

Em meu projeto de pesquisa, pretendo utilizar as respostas dos estudantes no teste cognitivo de língua portuguesa de avaliações em larga escala para investigar os seus erros e como eles podem ser utilizados para compreender o seu desempenho.

Em virtude do meu trabalho no Cebraspe, eu tenho acesso aos dados das avaliações realizadas pela instituição, entre as quais, a Prova/Provinha São Paulo realizada em 2018, com o devido sigilo de tais informações. Visando à consecução das análises do meu projeto de pesquisa, eu gostaria de solicitar a permissão do Núcleo Técnico de Avaliação da Secretaria Municipal de Educação de São Paulo para utilizar as respostas dos estudantes do 5º e do 9º ano no teste de língua portuguesa aplicado em 2018. Destaco que serão publicados no trabalho apenas os dados estatísticos e a interpretação dos resultados, sem comprometer o sigilo dos itens aplicados.

Tal análise pretende contribuir para uma compreensão mais ampla do desempenho dos estudantes da rede por meio da construção de uma escala que leve em consideração os erros que eles cometeram no teste. Esse trabalho será de grande contribuição também para o Cebraspe que, como instituição de pesquisa em avaliação, busca por meio desses estudos a melhoria das suas atividades. Me comprometo a, ao final da pesquisa, enviar o trabalho para a Secretaria Municipal de Educação de São Paulo, que poderá utilizar os resultados para a melhoria dos seus processos pedagógicos.

Agradeço a atenção e aguardo o seu retorno.

Atenciosamente

Laís S. Antonietto
+55 61 999521172

Carta de solicitação de acesso aos dados (conclusão)

Claudio Maroja <cmaroja@sme.prefeitura.sp.gov.br> Mon, Jul 20, 2020 at 5:20 PM
To: Laís Antonietto <lsantonietto@gmail.com>

Prezada Laís,

Parabéns pela iniciativa e ficamos felizes em colaborar com a pesquisa acadêmica e acrescento que, caso haja necessidade, mais dados podem ser obtidos no seguinte endereço http://dados.prefeitura.sp.gov.br/pt_PT/

Atenciosamente,

Claudio Maroja
SME/COPED/NTA - Núcleo Técnico de Avaliação
☎ 3396-0221 ✉ cmaroja@prefeitura.sp.gov.br
http://portal.sme.prefeitura.sp.gov.br

De: Laís Antonietto <lsantonietto@gmail.com>
Enviado: segunda-feira, 20 de julho de 2020 17:15
Para: Claudio Maroja <cmaroja@sme.prefeitura.sp.gov.br>
Assunto: Solicitação de autorização de acesso a dados - PROVA SÃO PAULO 2018

[Quoted text hidden]

Laís Antonietto <lsantonietto@gmail.com> Mon, Jul 20, 2020 at 5:26 PM
To: Claudio Maroja <cmaroja@sme.prefeitura.sp.gov.br>

Prezado Maroja,

Muito obrigada! Agradeço também a indicação do link.

Atenciosamente,

Laís S. Antonietto
+55 61 999521172

[Quoted text hidden]