



UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
DEPARTAMENTO DE MATEMÁTICA



ISABELA MARIA MELLE BAPTISTA

ESTUDO DE FERRAMENTAS ESTATÍSTICAS PARA A ANÁLISE DE ITENS EM RELATÓRIOS
PEDAGÓGICOS

SÃO CARLOS
2022

ISABELA MARIA MELLE BAPTISTA

ESTUDO DE FERRAMENTAS ESTATÍSTICAS PARA A ANÁLISE DE ITENS EM RELATÓRIOS
PEDAGÓGICOS

Monografia apresentada ao Curso de Licenciatura em Matemática para aprovação na disciplina de Trabalho de Conclusão de Curso.

Orientador: Prof. Dr. Jean Piton Gonçalves

SÃO CARLOS
2022



FUNDAÇÃO UNIVERSIDADE FEDERAL DE SÃO CARLOS
COORDENAÇÃO DOS CURSOS DE GRADUAÇÃO EM MATEMÁTICA - CCM/CCET
 Rod. Washington Luís km 235 - SP-310, s/n - Bairro Monjolinho, São Carlos/SP, CEP 13565-905
 Telefone: (16) 33518221 - <http://www.ufscar.br>

DP-TCC-FA nº 8/2022/CCM/CCET

Graduação: Defesa Pública de Trabalho de Conclusão de Curso
Folha Aprovação (GDP-TCC-FA)

FOLHA DE APROVAÇÃO

ISABELA MARIA MELLE BAPTISTA

ESTUDO DE FERRAMENTAS ESTATÍSTICAS PARA A ANÁLISE DE ITENS EM RELATÓRIOS PEDAGÓGICOS

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos – Campus São Carlos

São Carlos, 25 de abril de 2022

ASSINATURAS E CIÊNCIAS

Cargo/Função	Nome Completo
Orientador	Jean Piton Gonçalves
Membro da Banca 1	Sávio Brochini Rodrigues
Membro da Banca 2	Wladimir Seixas



Documento assinado eletronicamente por **Wladimir Seixas, Professor(a) do Magistério Superior**, em 17/05/2022, às 17:02, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Sávio Brochini Rodrigues, Professor(a) Adjunto(a)**, em 17/05/2022, às 18:15, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Jean Piton Goncalves, Professor(a) Adjunto(a)**, em 19/05/2022, às 08:04, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site <https://sei.ufscar.br/autenticacao>, informando o código verificador **0687281** e o código CRC **0E2D7FBE**.

Referência: Caso responda a este documento, indicar expressamente o Processo n°
23112.012371/2022-96

SEI n° 0687281

Modelo de Documento: Grad: Defesa TCC: Folha Aprovação, versão de 02/Agosto/2019

Dedico este trabalho aos meus pais, Neiva e Antonio.

AGRADECIMENTOS

Aos meus pais e irmãos, por todo incentivo e dedicação. Aos meus colegas de curso, Natalia Stefany Guirelli Molina, Juliana Sousa Paz, Gabriel Lacerda Ventura e Pedro Augusto Morelli, pelo companheirismo e por todos os momentos de descoberta e aprendizados. A todos os professores que contribuíram para a minha formação. Ao professor Dr. Jean Piton Gonçalves, pela orientação e apoio. À Universidade Federal de São Carlos (UFSCar) essencial no meu processo de formação educacional, profissional e como cidadã.

"Todos nós sabemos alguma coisa. Todos nós ignoramos alguma coisa.

Por isso, aprendemos sempre"

Paulo Freire ([FREIRE, 2017](#), p. 31).

RESUMO

A presente monografia aborda sobre a análise de itens de Matemática em Relatórios Pedagógicos em avaliações externas. Nessa direção, é apontada na literatura que professores de matemática apresentam dificuldades na interpretação e elaboração de tais relatórios. Visando preencher esta lacuna e em uma perspectiva quali-quantitativa, este trabalho aborda, do ponto de vista teórico, as principais ferramentas da avaliação educacional e da Teoria Clássica de Testes (TCT) presentes nos relatórios oficiais governamentais. Do ponto de vista prático, esta monografia interpretou o Relatório do SAEB 2017 no que se refere a análise de itens de matemática. Visando facilitar o cálculo de coeficientes estatísticos por parte de professores/gestores, foram implementadas planilhas eletrônicas dinâmicas que calculam determinados índices da TCT. No final, é conduzido um estudo de caso com itens de Matemática do ENEM da edição 2019, em que a partir de uma amostra aleatória de resposta daquela prova de Matemática, a planilha foi aplicada e alguns itens analisados.

Palavras-chave: Avaliação Educacional. Teoria Clássica de Testes. Relatórios do SAEB.

ABSTRACT

This degree final monograph the analysis of math items in pedagogical reports of external evaluations. In this direction, it is pointed out in the literature that math teachers have difficulties in interpreting and in making such reports. Aiming to fill this gap and in a qualitative-quantitative perspective, this work approaches, from a theoretical point of view, the main tools of educational assessment and the Classical Test Theory (TCT) present in official government reports. From a practical point of view, this work interpreted the 2017 SAEB Report with regard to the analysis of math items. In order to facilitate the calculation of statistical coefficients by teachers/managers, dynamic electronic spreadsheets that calculate certain TCT indices were implemented. Last, a case study is carried out with ENEM Mathematics items from the 2019 edition, in which, from a random sample of responses, the worksheet was applied and some items analyzed.

Keywords: Educational Assessment. Classical Test Theory. SAEB reports.

LISTA DE FIGURAS

Figura 2.1 – FUVEST 2021. 1ª fase - Conhecimentos Gerais - Questão 06 (matemática)	15
Figura 5.1 – Coeficientes estatísticos clássicos de um determinado item.	28
Figura 5.2 – Item 22227 - Matemática - 5º ano do Ensino Fundamental - SAEB 2017	30
Figura 5.3 – Item 34762 - Matemática - 9º ano do Ensino Fundamental - SAEB 2017	31
Figura 5.4 – Código 29036 - Matemática - 3º série do Ensino Médio - SAEB 2017	32
Figura 6.1 – Recorte da planilha interativa no <i>Google Sheets</i>	37
Figura 6.2 – Cálculo das estatísticas para interpretação dos itens no <i>Google Sheets</i>	37
Figura 6.3 – Análise do vetor-resposta no <i>Google Sheets</i>	39
Figura 6.4 – Frequência na escolha das alternativas do item <i>Google Sheets</i>	40
Figura 6.5 – Cálculo da média de acertos em cada alternativa no item <i>Google Sheets</i>	41
Figura 6.6 – Cálculo do coeficiente bisserial para cada alternativa <i>Google Sheets</i>	41
Figura 6.7 – Item 148 — Matemática — ENEM 2019 — Caderno Amarelo	42
Figura 6.8 – Item 165 — Matemática — ENEM 2019 — Caderno Amarelo	43

LISTA DE TABELAS

Tabela 5.1 – Informações gerais do item - 22227	31
Tabela 5.2 – Fonte: Adaptada pela autora	31
Tabela 5.3 – Informações gerais do item - 34762	32
Tabela 5.4 – Fonte: Adaptada pela autora	32
Tabela 5.5 – Informações gerais do item - 34762	33
Tabela 5.6 – Coeficientes estatísticos clássicos dos itens Saeb 2017 – Matemática – Ensino Fundamental I e II	33
Tabela 5.7 – Tabela adaptada pela autora com base no Relatório do SAEB - 2017	33
Tabela 5.8 – Coeficientes estatísticos clássicos dos itens Saeb – Matemática – 2017	35
Tabela 5.9 – Tabela adaptada pela autora com base no Relatório do SAEB - 2017	35
Tabela 6.1 – Informações gerais do item — 148	42
Tabela 6.2 – Informações gerais do item - 165	43
Tabela 6.3 – Resultados quantitativos do estudo de caso do ENEM 2019 – Matemática – Caderno amarelo	44
Tabela 6.4 – Adaptada pela autora	44

SUMÁRIO

1	INTRODUÇÃO	10
2	AVALIAÇÃO E TESTES EDUCACIONAIS	12
2.1	CLASSES DE AVALIAÇÃO EDUCACIONAL	12
2.2	TESTES EDUCACIONAIS	14
2.3	SAEB	16
3	ESTATÍSTICA PARA ANÁLISE DE ITENS	17
3.1	TIPOS DE DADOS	17
3.2	MEDIDAS DE POSIÇÃO E DE DISPERSÃO	18
3.3	CORRELAÇÃO BISSERIAL	20
3.4	COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON	21
3.5	COEFICIENTE DE CORRELAÇÃO DE SPEARMAN-BROWN	22
4	TEORIA CLÁSSICA DE TESTES	23
4.1	ÍNDICE DE DIFICULDADE	24
4.2	ÍNDICE DE DISCRIMINAÇÃO	25
4.3	FIDEDIGNIDADE E VALIDADE DE UM TESTE	25
5	RESULTADOS 1	27
5.1	OS RELATÓRIOS DO SAEB	27
5.2	INTERPRETAÇÃO DA ANÁLISE DE ITENS DE MATEMÁTICA NO RELATÓRIO PEDAGÓGICO DO SAEB - 2017	30
5.3	INTERPRETAÇÃO GERAL	30
6	RESULTADO 2	36
6.1	A PLANILHA	36
6.2	ESTUDO DE CASO	41
7	CONCLUSÃO	45

1 INTRODUÇÃO

O ato de avaliar está sempre atrelado à ideia de julgamento e atribuição de valor. Pontua-se que a avaliação deve ser entendida como sendo uma atividade formal de maneira que provas ou testes, servem como instrumentos avaliativos coletando evidências que permitem ao avaliador afirmar se os objetivos esperados foram alcançados ou não.

No contexto da educação brasileira, uma importante categoria de avaliação são as avaliações externas, que contribuem com determinadas informações que norteiam a qualidade do ensino. Tais avaliações são realizadas em diferentes níveis (municipal, estadual e nacional) e desempenham um papel relevante no desenho das políticas públicas. Nesse contexto citamos o Exame Nacional do Ensino Médio (ENEM) ¹ e Sistema de Avaliação da Educação Básica (SAEB).

Nas concepções de avaliação educacional, os testes de múltipla escolha constituem-se como um instrumento avaliativo de grande eficiência e versatilidade. Além disso, destaca-se que essa modalidade de instrumentação pode ser analisada estatisticamente. Essas estatísticas aparecem em relatórios pedagógicos, cuja finalidade é apresentar dados e informações referentes às avaliações. As informações obtidas por meio das 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” (SOUSA; FERREIRA, 2019b, p.16). Destaca-se, que parte das avaliações externas nacionais utilizam-se de metodologias estatísticas e psicométricas, como, por exemplo, Teoria Clássica de Testes (TCT) que, essencialmente, busca medir e validar sistematicamente as competências e/ou habilidades do estudante e/ou examinado em um teste avaliativo.

Um exemplo de Relatório Pedagógico que é disponibilizado às escolas é do SAEB, principal ferramenta de estudo deste trabalho. Os estudos de Silva e Darsie (2008) apontam “acerca das dificuldades apresentadas pelos professores de matemática na elaboração dos relatórios avaliativos de seus alunos”(SILVA; DARSIE, 2008, p.16). Uma vez que há a necessidade, dos gestores e professores, em compreender as estatísticas e coeficientes quanto à análise de itens das avaliações externas e que eu, como estudante do curso de licenciatura em Matemática e futura docente do ensino básico, tenho por motivação nesta Monografia, interpretar os índices e coeficientes estatísticos disponibilizados no Relatório Pedagógico do Saeb (2017), tendo como foco os itens de matemática e aplicar a TCT em dados educacionais simulados por meio do uso de tecnologias da informação, permitindo que gestores e professores possam aplicar tal teoria em escolas e/ou instituições de ensino. Para isso, o Capítulo 2 abordará “Avaliação e Testes Educacionais” visando apresentar as classes de avaliação educacional, os tipos de testes educacionais e compreensão da estrutura do Saeb. O Capítulo 3, terá como foco uma breve revisão

¹ O Exame Nacional do Ensino Médio (ENEM) tem como objetivo de avaliar o desempenho escolar dos estudantes ao término da educação básica. Em 2009, o exame aperfeiçoou sua metodologia e passou a ser utilizado como mecanismo de acesso à educação superior. (INEP, 2021a)

acerca da estatística para análise de itens. O Capítulo 4, tem como foco a discussão acerca da teoria clássica de testes. Serão estudados os índices de dificuldade, índice de discriminação e fidedignidade e validade de um teste de múltipla escolha. O Capítulo 5 apresentará os resultados referente a interpretação da análise de itens de matemática no relatório pedagógico do Saeb -2017 e o Capítulo 6 apresentará os cálculos envolvidos na TCT feitos por meio de um Banco de Dados gerados por simulação e a análise dos resultados obtidos. Já o Capítulo 7 traz as Considerações Finais deste TCC.

2 AVALIAÇÃO E TESTES EDUCACIONAIS

Avaliar é algo inerente ao ser humano. Fazemos avaliações o tempo todo, em diversas situações. De acordo com [Marinho-Araujo e Rabelo \(2015, p.444\)](#) este é um processo “que envolve concepções, crenças, valores, princípios, teorias, conceitos, metas, desejos, trajetórias”. Para [HAYDT \(2011, p.290\)](#) o conceito de avaliar, pode ser entendido como:

[...] julgar ou fazer uma apreciação sobre alguém ou alguma coisa, tendo como base uma escala de valores. Assim, a avaliação consiste na coleta de dados quantitativos e qualitativos e na interpretação desses dados com base em critérios previamente definidos.

Assim, diferentemente dos atos avaliativos informais que emitimos em nosso cotidiano, a avaliação deve ser compreendida como sendo uma prática sistematizada, organizada e formal, e que “é feita de acordo com os objetivos escolares implícitos ou explícitos, que, por sua vez, refletem valores e normas sociais” ([CHUIEIRE, 2008, p.51](#)).

No contexto escolar, o termo avaliação é frequentemente associado a prova, nota ou classificação. Estas se relacionam com o ato de avaliar, mas não são de fato uma avaliação. “A atividade educativa não tem por meta atribuir notas, mas realizar uma série de objetivos que se traduzem em termos de mudanças de comportamento dos alunos” ([HAYDT, 2011, p.9](#)). Assim “a avaliação na educação deve ser entendida, portanto, como um processo amplo, com desdobramentos coletivos e institucionais, além de individuais” ([MARINHO-ARAUJO; RABELO, 2015, p.444](#)).

2.1 CLASSES DE AVALIAÇÃO EDUCACIONAL

Na perspectiva de avaliações orientadas para o aperfeiçoamento de práticas pedagógicas, a avaliação pode: diagnosticar, controlar e classificar. Nesse contexto, podemos contemplar três classificações da avaliação educacional:

Avaliação diagnóstica: tem como objetivos básicos a identificação das competências, habilidades e conhecimentos que o examinado já possui. Trata-se da avaliação:

[...] realizada no início de um curso, período letivo ou unidade de ensino, com a intenção de constatar se os alunos apresentam ou não o domínio dos pré-requisitos necessários, isto é, se possuem os conhecimentos e habilidades imprescindíveis para as novas aprendizagens. É também utilizada para caracterizar eventuais problemas de aprendizagem e identificar suas possíveis causas, numa tentativa de saná-los ([HAYDT, 2011, p.16](#))

[Miquelante et al. \(2017\)](#) pontua que, com o diagnóstico obtido nessa modalidade de avaliação, o professor tem informações que o ajudam a identificar o conhecimento prévio dos examinados a respeito dos conteúdos do curso ou nível escolar e, a partir desta etapa, adaptar o projeto pedagógico a fim de garantir a superação das dificuldades encontradas.

Avaliação formativa: visa avaliar o que foi ensinado. Para [Miquelante et al. \(2017, p.269\)](#), essa categoria de avaliação promove a interação entre o professor e o aluno durante todo o processo do ensino e aprendizagem “uma vez que auxilia os envolvidos com informações acerca dos objetivos alcançados e os esforços necessários para desenvolver o que ainda não foi atingido”.

[Piton-Gonçalves \(2012, p.10\)](#) destaca que:

A avaliação formativa está muito ligada ao mecanismo de *feedback*, à medida que também permite ao professor detectar e identificar deficiências na forma de ensinar, possibilitando reformulações no seu trabalho didático, visando aperfeiçoá-lo.

Bem como pontua [Taras \(2010, p.127\)](#), essa categoria de avaliação centraliza significados, técnicas e procedimentos com o intuito de auxiliar a aprendizagem por meio do *feedback*, sendo este meio um aspecto fundamental da avaliação formativa.

Avaliação somativa: tem como pretensão aferir os resultados de aprendizagem alcançado pelo examinado, conforme os objetivos que foram pré-estabelecidos. É preferencialmente realizada após uma sequência de estudos, ou então ao final de um determinado ciclo, sendo desvinculada do processo de ensino. Além disso, essa modalidade de avaliação, tem como características classificar segundo algum critério estabelecido ou ainda promover o indivíduo de um nível para outro.

[...] com função classificatória, realiza-se ao final de um curso, período letivo ou unidade de ensino, e consiste em classificar os alunos de acordo com os níveis de aproveitamento previamente estabelecidos, geralmente tendo em vista sua promoção de uma série para a outra, ou de um grau para outro ([HAYDT, 2011, p.18](#))

Ressalta-se que “esses três tipos de avaliações estão intimamente vinculadas para garantir a eficiência do sistema de avaliação e a eficácia do processo ensino-aprendizagem. Elas são baseadas e definidas em objetivos educacionais claros” ([HAYDT, 2011, p.18](#)).

No contexto da avaliação somativa, a Avaliação externa visa obter um diagnóstico do desempenho a fim de financiar e acompanhar políticas públicas. O planejamento desta avaliação, normalmente é feito por órgãos especializados a fim de proporcionar uma estrutura adequada a esta categoria. Além disso, a avaliação externa tem como principal instrumento avaliativo o uso de testes, geralmente os testes de múltipla escolha, tema principal deste trabalho e que será melhor discutido na próxima seção. Para [Machado \(2017, p.71\)](#) a avaliação externa é definida como sendo todo procedimento avaliativo da performance das escolas desencadeado e operacionalizado por agentes externos ao cotidiano escolar. Além disso, dependendo dos arranjos do processo desta categoria de avaliação, pode haver a participação de profissionais das escolas examinadas, porém a decisão da implementação de uma avaliação do desempenho das escolas é sempre externas a elas.

Destaca-se que as avaliações externas, quando direcionadas para um grande número de examinados que excede a quantidade de alunos de uma turma ou até mesmo de uma escola, é denominada *avaliação em larga escala*, que surge da ideia da promoção da qualidade do ensino e de identificar a necessidade de implementação de políticas públicas educacionais, e são realizadas em diferentes níveis (municipais, estaduais e federais).

Sousa e Ferreira (2019a, p.16) destacam a importância das avaliações em larga escala e de seus indicadores de qualidade ¹

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 sócio econômico da clientela escolar é decisivo no desempenho acadêmico individual

No contexto das avaliações externas brasileiras de grande relevância no cenário educacional citamos o Exame Nacional do Ensino Médio (ENEM) ² Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo (SARESP) ³ e Sistema de Avaliação da Educação Básica (Saeb), que será melhor discutido posteriormente.

2.2 TESTES EDUCACIONAIS

No âmbito educacional, umas das ferramentas mais utilizadas para avaliar, são os testes que, segundo HAYDT (2011, p.10) têm por objetivo "verificar um desempenho através de situações previamente organizadas". É possível extrair medidas e resultados, de modo a estipular objetivos e metas para o processo de ensino e aprendizagem. Destaca-se que um teste "pode ser realizado direta ou indiretamente, por um especialista ou por um ambiente educacional (PITON-GONÇALVES, 2012, p.9).

De modo geral, um teste educacional tem em sua composição elementos, denominados itens. HAYDT (2011) destaca alguns modelos de tipos de itens a serem inseridos na elaboração de um teste, tais como, itens objetivos, de respostas curtas, de lacuna, de certo e errado, de associação, correlação ou combinação e de múltipla escolha. Além disso, a escolha do tipo de item a ser eleito para elaboração do teste, irá depender dos objetivos a serem analisados pelo aplicador.

Geralmente em testes educacionais em larga escala, o formato de resposta aos itens seguem duas abordagens: dissertativa e múltipla escolha.

¹ Para saber mais: Estudos em Avaliação Educacional. vol.2 no.66 São Paulo set./dez.2016. Disponível em: <<http://educa.fcc.org.br/scielo>>. Acesso em setembro de 2021

² O Exame Nacional do Ensino Médio (ENEM) tem como objetivo de avaliar o desempenho escolar dos estudantes ao término da educação básica. Em 2009, o exame aperfeiçoou sua metodologia e passou a ser utilizado como mecanismo de acesso à educação superior (INEP, 2021a).

³ O Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo-(SARESP) é uma avaliação voltada a alguns anos do Ensino Fundamental e ao terceiro ano do Ensino Médio, na qual tem a finalidade de produzir um diagnóstico da situação da escolaridade básica do Estado de São Paulo, visando orientar os gestores do ensino no monitoramento das políticas voltadas para a melhoria da qualidade educacional (FDE, 2021).

Em itens com respostas dissertativas, o examinado discorre e apresenta argumentos, por meio da escrita. A correção de um item dissertativo demanda especialistas de conteúdo, aumentando os custos de sua correção e ampliando o tempo *feedback* do teste para os examinados.

Visando minimizar os (i) custos de aplicação e correção e (ii) priorizar respostas objetivas, um item de múltipla escolha é composto, fundamentalmente, por um texto-base (texto inicial), enunciado (é a pergunta em si) e as opções de resposta (chamadas também de alternativas, são as respostas pré determinadas). Um formato de opção de resposta muito utilizado em avaliações em larga escala é o dicotômico, em que dentre as opções, apenas uma é a correta e as demais são distratoras.

A Figura 2.1 traz um exemplo de um item de múltipla escolha, referente ao conteúdo de matemática, da prova da Fundação Universitária para o Vestibular (FUVEST)⁴, 1ª fase, realizada no ano de 2021, cuja a alternativa correta é a letra "(B)".

Figura 2.1 – FUVEST 2021. 1ª fase - Conhecimentos Gerais - Questão 06 (matemática)

06

Uma treinadora de basquete aplica o seguinte sistema de pontuação em seus treinos de arremesso à cesta: cada jogadora recebe 5 pontos por arremesso acertado e perde 2 pontos por arremesso errado. Ao fim de 50 arremessos, uma das jogadoras contabilizou 124 pontos. Qual é a diferença entre as quantidades de arremessos acertados e errados dessa jogadora?

(A) 12
(B) 14
(C) 16
(D) 18
(E) 20

Fonte: Fuvest (2021)

Para Osterlind (1998, p.162), o formato de múltipla escolha para itens, dentre suas principais vantagens, é o de ser mais flexível para avaliar uma diversidade de conteúdos e processos psicológicos, se comparado a outro formato de itens. Outro benefício deste instrumento é que ele possibilita uma interpretação precisa que pode levar a importantes evidências para a validade do teste relacionado ao conteúdo.

Neste trabalho, teremos como foco de estudo os testes com itens de múltipla escolha, em avaliações nacionais de larga escala.

⁴ Criada pela Universidade de São Paulo (USP), a Fundação Universitária para o Vestibular (FUVEST) foi criada para realizar o exame vestibular da USP, de acordo com as diretrizes estabelecidas pelo Conselho de Graduação da Universidade. Para saber mais, acesse: <<https://www.fuvest.br/sobre/>>

2.3 SAEB

O SAEB⁵ é um conjunto de avaliações externa em larga escala, proposta pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep)⁶ com relevância no cenário educacional e que tem por objetivo obter um diagnóstico da educação básica brasileira e também a identificar fatores que podem interferir no desempenho do estudante (INEP, 2021c). Além disso, é aplicado por meio de testes e questionários que são realizados em larga escala tanto para a rede pública de ensino quanto para a rede privada.

Até o ano de 2020 o SAEB era realizado a cada dois anos para alunos dos 5º e 9º anos do ensino fundamental e também para alunos do 3º ano do ensino médio abrangendo Língua Portuguesa e Matemática, como sendo as áreas do conhecimento avaliadas. No dia 11 de fevereiro de 2021, no Diário Oficial da União (DOU), o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira publicou a portaria nº 10⁷, estabelecendo "parâmetros e diretrizes gerais para implementação do novo Sistema de Avaliação da Educação Básica (Saeb), no âmbito da Política Nacional de Avaliação da Educação Básica "(INEP, 2021b).

Com as eventuais mudanças, o exame passará a ocorrer anualmente a partir de 2021 e será aplicada para todos os anos e séries a partir do do 2º ano do ensino fundamental. Além disso, tanto os alunos das redes públicas de ensino quanto alunos das redes privadas serão avaliados. A mudança na frequência das avaliações visa obter informações com maior precisão por unidade escolar e, à partir disso fazer intervenções pedagógicas em um espaço de tempo menor.

O INEP emite Relatórios Pedagógicos e/ou Técnicos no que se refere aos resultados da aplicação do SAEB e que acabam sendo utilizados por gestores e professores em no ambiente escolar. Um dos componentes desses relatórios é a análise estatística de algumas questões da prova. Os Capítulos 3 e 4 versam sobre as principais estatísticas e teorias envolvidas.

⁵ <<https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb>>

⁶ É uma autarquia federal vinculada ao Ministério da Educação (MEC). Sua missão é subsidiar a formulação de políticas educacionais dos diferentes níveis de governo com intuito de contribuir para o desenvolvimento econômico e social do país. (Para mais informações consultar: <http://portal.inep.gov.br/conheca-o-inep>)

⁷ Para saber mais, acesse: <https://www.gov.br/inep/pt-br/assuntos/noticias/saeb/portaria-estabelece-diretrizes-para-o-novo-saeb>

3 ESTATÍSTICA PARA ANÁLISE DE ITENS

A estatística pode ser compreendida como sendo um conjunto de métodos matemáticos e que permite a organização de dados e informações. Além de grande aplicabilidade em pesquisas científicas, esta área também se aplica a diversos segmentos, bem como, pesquisas eleitorais, demográficas, etc.

No contexto educacional, um dos campos de grande aplicação da estatística, é nos processos avaliativos, em particular nos instrumentos avaliativos, como a prova, por exemplo. Esta ferramenta pode ser utilizada desde a elaboração, correção, análise e até a atribuição de notas, em qualquer disciplina. Visto que este trabalho visa interpretar a análise de itens em relatórios pedagógicos, abordaremos neste capítulo algumas ferramentas estatísticas que irão auxiliar neste processo.

3.1 TIPOS DE DADOS

Os dados são objetos fundamentais da estatística, os quais podem ser obtidos por meio (i) de uma população, que é o conjunto de todos os elementos, no qual deseja-se entender características específicas que está sob investigação, e a (ii) uma amostra, que é um subconjunto qualquer do conjunto população.

A característica de interesse, que é medida nos elementos da população ou amostra, podendo variar e assumir valores numéricos ou não numéricos, é denominada de variável. Estas podem ser classificadas em :

Qualitativa: assume como possíveis valores qualidades ou atributos. Segundo [Morettin e Bussab \(2017\)](#) são definidas como variáveis de possíveis realizações (qualidade) ou atributo do indivíduo pesquisado. Além disso, são subdividas em dois grupos: nominal e ordinais.

Quantitativas: são variáveis que podem ser medidas em uma escala quantitativa, isto é, apresentam valores numéricos que fazem sentido e são subdividas em: discretas, contínuas e dicotômicas.

Qualitativas nominais: possuem dados não mensuráveis, são categorizados sem nenhuma ordenação nas possíveis realizações. De modo geral, diz-se que não há ordenação nas possíveis respostas. Exemplo: estado civil, fumante ou não fumante, etc.

Qualitativas ordinais: possuem dados não mensuráveis, são categorizados sob uma ordem em seus resultados. Diz-se que há relação de ordem nas possíveis respostas. Exemplo: grau de escolaridade.

Quantitativas discretas: possuem dados mensuráveis podendo assumir apenas um número finito enumerável, são variáveis de contagem. Exemplo: número de filhos, números de aparelhos eletrônicos, etc.

Quantitativas contínuas: os possíveis valores pertencem a um intervalo aberto ou fechado, dos números reais. Exemplos: peso de indivíduo, altura, salário, pressão arterial, etc.

Quantitativas dicotômicas: possuem dados mensuráveis binários para receber a qualificação. São casos em que não existe uma terceira opção para ser designado. Exemplos: acertar ou errar um item de uma prova, resultado de um lançamento de moeda, etc.

3.2 MEDIDAS DE POSIÇÃO E DE DISPERSÃO

As medidas de posição, tem por objetivo resumir um conjunto de dados, apresentando um ou alguns valores que sejam representativos da série toda (MORETTIN; BUSSAB, 2017). Há três tipos principais de medidas de posição para dados discretos, são elas:

Média aritmética simples: consiste na soma dos valores obtidos em uma observação dividida pela número delas. Sendo expressa por:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} \quad (3.1)$$

tal que $x_1 + x_2 + x_3 + \dots + x_n$ são os n valores (distintos ou não) da variável X e n o número de observações

Mediana: “é a realização que ocupa a posição central da série de observações, quando estão ordenadas em ordem crescente” (MORETTIN; BUSSAB, 2017, p.35), isto é, a mediana é uma separatriz que divide o conjunto em duas partes partes iguais. Considera-se as observações ordenadas abaixo em ordem decrescente,

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n-1)} \leq x_{(n)} \quad (3.2)$$

sendo $x_{(1)}$ a menor observação, a segunda $x_{(2)}$ e assim sucessivamente. As observações ordenadas acima são denominadas *estatísticas de ordem*.

De maneira formal, a mediana da variável X pode ser definida como:

Se n é ímpar, a mediana é dada por $md(X) = X_{(\frac{n+1}{2})}$.

Se n é par, a mediana é dada por $md(X) = \frac{X_{(\frac{n+1}{2})} + X_{(\frac{n}{2}+1)}}{2}$

Para o cálculo da posição da mediana, tem-se:

$$\text{Posição da Mediana} = \frac{N + 1}{2} \quad (3.3)$$

Sendo N o número total de observações.

Moda: é classificada como sendo a realização com maior frequência em um conjunto de valores observados. Além disso, é subdividida em duas classes: multimodal ¹ e amodal ². Sua aplicação se restringe apenas para mostrar qual é a observação com maior repetição no conjunto de dados, sendo de uso pouco habitual.

Para uma visualização mais adequada das medidas apresentadas acima, considere a seguinte situação: foram coletadas e adaptadas ³ as notas de classificação de 13 candidatos ingressantes no curso de Matemática, do período noturno, na Universidade Federal de São Carlos (UFScar), no ano de 2021 ⁴, sendo elas: 753, 740, 737, 733, 733, 731, 725, 723, 722, 721, 719, 712, 710. A média simples é dada por:

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i = \bar{x} = \frac{1}{13} \sum_{i=1}^{13} x_i = \frac{x_1 + x_2 + \dots + x_{13}}{13} \\ &= \frac{753 + 740 + 737 + 733 + 733 + 731 + 725 + 723 + 722 + 721 + 719 + 712 + 710}{13} = \\ &= 728 \end{aligned}$$

No caso da moda, nota-se que o valor que ocorre com maior frequência é o 733 que aparece duas vezes. Para o caso da mediana, o primeiro passo é organizar os dados em ordem crescente e posteriormente utilizar a definição de mediana, apresentada anteriormente. Visto que a quantidade de dados coletados é ímpar, a *mediana* é dada por:

$$X_{\frac{13+1}{2}} = X_{\frac{14}{2}} = X_7$$

¹ conjunto de dados com mais de uma moda

² conjunto de dados em que não há um valor predominante.

³ As notas foram arredondadas para o seus inteiros, sendo descartada a parte decimal.

⁴ Notas obtidas através do site <<https://sisu.mec.gov.br///selecionados>>

Ou seja, a mediana está localizada na sétima posição. Além disso, pela Equação (3.3) pode-se verificar a posição da mediana, é dado por:

$$\frac{N + 1}{2} = \frac{13 + 1}{2} = \frac{14}{2} = 7$$

Sendo assim, a nota 725 que está disposto na sétima posição, é a mediana procurada.

As informações obtidas pelas medidas de posição não exprimem quanto à dispersão do conjunto de dados. Para isso, as medidas de dispersão registram o grau de variabilidade dos dados observados em torno da média aritmética, sendo a variabilidade de um conjunto a maior ou a menor diferença dos dados de uma variável em torno de um valor de tendência central como ponto de comparação. Usualmente, as duas medidas de posição mais utilizadas são:

Variância: considera os extremos e também os intermediários, além disso, relaciona os desvios em torno da média. É dada pela seguinte expressão

$$\text{var}(X) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} \quad (3.4)$$

Desvio Padrão: O desvio padrão é simplesmente a raiz quadrada da variância. Ressalta-se que esta é uma medida muito conveniente para a análise da dispersão de um conjunto de dados, proporcionando a confiabilidade de um conjunto de medidas. Formalmente, tem-se que o desvio padrão é dado por:

$$Dp(X) = \sqrt{\text{var}(X)} \quad (3.5)$$

3.3 CORRELAÇÃO BISSERIAL

O *Coefficiente de Correlação Bisserial* é utilizado para medir a correlação entre duas variáveis contínuas (itens, por exemplo) porém uma delas se dicotomizou (assume apenas dois valores). Bem como pontua Soares (2005, p.85), a correlação bisserial e a correlação ponto bisserial pode ser entendida da seguinte forma:

Considere o caso de testes constituídos por itens binários ou dicotômicos, isto é, itens para os quais se admite duas respostas possíveis. A correlação bisserial e a correlação ponto bisserial são medidas estatísticas que medem a correlação do resultado de um item em particular do teste com o resultado do teste (isto é, o escore bruto total), sendo, portanto, uma medida da capacidade de discriminação do item em relação ao resultado do teste.

De maneira formal, a correlação ponto bisserial é dada por:

$$r_{b,i} = \frac{\bar{X}_{i1} - \bar{X}}{Dp(X)} \cdot \frac{\sqrt{p_i}}{\sqrt{1 - p_i}}, -1 \leq r_{b,i} \leq 1 \quad (3.6)$$

Tal que \bar{X}_{i1} é o escore médio dentre os indivíduos que acertaram o item; \bar{X} é o escore médio geral; S o desvio padrão da medida de desempenho no teste para todos os indivíduos (escore); p_i medida de proporção de acertos no item e $1 - p_i$ a proporção de erros no item.

Um exemplo de aplicação das correlações bisseriais é na prova do SAEB, é através destas correlações que são identificados itens que possuem algum problema no gabarito.

(...) de acordo com os procedimentos adotados no SAEB no ano de 2005, os itens com coeficiente bisserial do gabarito menor ou igual a 0,15; itens com dois coeficientes bisseriais de distratores (alternativas erradas) maiores que 0,10 ou coeficiente bisserial de um distrator maior que a bisserial da alternativa correta são encaminhados para a análise pedagógica (INEP, 2005). (ANDRADE; LAROS; GOUVEIA, 2010, p.426)

Posteriormente, após um grupo de especialistas realizar uma análise pedagógica, os itens apontados como problemáticos podem ter seu gabarito corrigido ou então ser abandonado nas análises adiante.

3.4 COEFICIENTE DE CORRELAÇÃO LINEAR DE PEARSON

Na análise de dados, é natural o estudo de mais de uma variável simultaneamente. Um ponto chave durante a construção de uma distribuição conjunto de duas variáveis qualitativas é caracterizar a combinação entre elas, ou seja, deseja-se identificar qual o grau de dependência entre elas, de maneira que seja possível prever o melhor resultado de uma quando é conhecido a realização da outra. Por exemplo, se deseja-se estimar qual a renda média de uma família residente na cidade de São Carlos, a informação complementar sobre a classe social a que ela pertence permite uma estimativa com um alto grau de precisão a renda, pois é sabe-se que há uma dependência entre as duas variáveis: renda familiar e classe social (exemplo adaptado Morettin e Bussab (2017, p.73)).

Sendo assim, destaca-se que a quantificação do grau de associação entre duas variáveis é realizada por meio do coeficientes de associação ou correlação. Tais medidas descrevem, através de um único número, o grau de dependência entre duas variáveis. Bem como pontua Morettin e Bussab (2017, p.76) "para maior facilidade de compreensão, esses coeficientes usualmente variam entre 0 e 1, ou entre -1 e $+1$, e a proximidade de zero indica falta de associação".

A equação do coeficiente de correlação linear de Pearson é dada por:

$$r = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \sqrt{n \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2}} \quad (3.7)$$

Sendo n o número de observações e $r \in [-1, 1]$.

3.5 COEFICIENTE DE CORRELAÇÃO DE SPEARMAN-BROWN

A correlação de Spearman-Brown ou coeficiente de correlação de postos de Spearman, é descrita como sendo uma medida não paramétrica ⁵ de correlação de postos, isto é, a dependência estatística entre a classificação de duas variáveis. Tal coeficiente analisa o grau de intensidade em que a relação entre duas variáveis pode ser expressa pelo uso de uma função monótona. De modo geral, a correlação de Spearman-Brown é igual ao coeficiente de Pearson, sendo aplicado aos valores da ordem de duas variáveis. Destaca-se que enquanto a correlação de Pearson mede as relações lineares, a correlação de Spearman-Brown avalia as relações monótonas, isto é, lineares ou não lineares.

O coeficiente de Spearman é adequado tanto para o caso de variáveis contínuas, como para o caso de variáveis discretas, incluindo variáveis ordinais ⁶. Esse coeficiente pode ser calculado pela seguinte fórmula:

$$r_s = 1 - \frac{6 \sum_{i=1}^n D^2}{n \cdot (n^2 - 1)} \quad (3.8)$$

Sendo, r_s a correlação de postos, D a diferença entre postos (relativa ao mesmo sujeito em ambas as variáveis) e n o número de observações.

⁵ Para mais informações, acesse : <<https://www.lume.ufrgs.br/bitstream/handle/10183/72818/000886232.pdf;sequence=1>>

⁶ Existe uma ordenação entre as categorias, por exemplo, estágio de doença, grau de escolaridade

4 TEORIA CLÁSSICA DE TESTES

A psicologia é uma ciência que preocupa-se com o estudo e compreensão do ser humano em um nível mental, por exemplo, inteligência, personalidade, memória, aprendizado, etc. Por ser um campo da ciência, a psicologia utiliza de instrumentos e métodos reconhecidos cientificamente, para avaliação de tais constructos. Uma das áreas que auxilia na construção de instrumentos de medidas é a Psicometria, sendo uma vertente da estatística, na psicologia, que fornece teorias para a construção de tais constructos “Etimologicamente, psicometria representa a teoria e a técnica de medida dos processos mentais, especialmente aplicada na área da Psicologia e da Educação” (PASQUALI, 2009a, p.993). De acordo com Primi (2012, p.300) a Psicometria é um campo particular da psicologia e designada “ao desenvolvimento de técnicas de mensuração de variáveis psicológicas, introduzindo um instrumental estatístico adequado as suas complexidades”.

Um das vertentes da psicometria, é a Teoria Clássica de Testes (TCT). De acordo com Sartes e Souza-Formigoni (2013, p.242), a TCT abrange um conjunto de conceitos e técnicas que tem sido aplicada “como base para o desenvolvimento de numerosos instrumentos de medidas e como ponto de referência para abordagens de medidas psicométricas”. De modo geral, a Teoria Clássica de Testes leva em consideração a pontuação total de um teste (escore), isto é, a soma das respostas obtidas em uma série de itens propostos. O modelo matemático da TCT, foi elaborado por Spearman (PASQUALI, 2009a) e aprofundado por Gulliksen (GULLIKSEN, 2013) dado pela seguinte expressão:

$$T = V + E \quad (4.1)$$

tal que, T é o escore bruto ou empírico do sujeito, que a soma dos pontos obtidos no teste, V o escore verdadeiro, que seria a magnitude real daquilo que o teste quer medir no sujeito e que seria o próprio T se não houvesse o erro da medida e E o erro cometido nesta medida (PASQUALI, 2009a)[p.69].

Assim, o escore empírico é a soma do escore verdadeiro e do erro. Além disso, a soma do escore verdadeiro e do erro com base nesse cálculo é obtido as informações referente ao sujeito que está sendo avaliado em comparação com toda a amostra. Bem como pontua Pasquali (2009a) o grande foco da TCT consiste em desenvolver formas (estatísticas) para controlar ou avaliar a relevância do erro cometido na medida (E). Campbell e Stanley (2015) detalham que os erros são decorrentes de uma série de fatores estranhos, bem como, problemas do próprio teste, estereótipos do sujeito, fatores históricos e ambientais aleatórios.

Destaca-se que para a avaliação de um teste dentro da TCT, há dois requisitos fundamentais, são eles: validade e fidedignidade. Tais temas serão discutidos de maneira mais aprofundada nas seções posteriores.

No contexto educacional, a Teoria Clássica de Testes é uma ferramenta que fornece ao

professor parâmetro para determinar a qualidade dos itens em uma avaliação, quantificando a aptidão de cada item em diagnosticar a proficiência do aluno examinado na habilidade avaliativa.

4.1 ÍNDICE DE DIFICULDADE

Na Teoria Clássica de Testes, o índice de dificuldade (ID) de um item é obtido por meio da quantidade de indivíduos que responderam corretamente ao item, no caso de itens dicotômicos (SARTES; SOUZA-FORMIGONI, 2013). Para o caso de itens politômicos ¹, o grau de dificuldade é determinado pela proporção de respostas a uma classe de escolha ou pela média das respostas de todos os indivíduos (SARTES; SOUZA-FORMIGONI, 2013; DEVELLIS, 2006; GRÉGOIRE; LAVEAULT, 2002).

Segundo Figueiredo et al. (2008, p.586) “a razão principal para medir a dificuldade de itens é a de descartar aqueles que possuem um nível impróprio de dificuldade, não informando diferenças individuais nem proporcionando variabilidade aos resultados de um teste”.

O índice de dificuldade (ID) pode ser calculado pela seguinte fórmula:

$$ID_i = \frac{A}{N} \quad (4.2)$$

Onde A é o número de sujeitos que acertaram o item, N o número total de sujeitos que responderam o item e $0 \leq ID_i \leq 1$. Assim, quanto menor o índice, maior a dificuldade. Sendo $ID_i = 1$ significa que todos os sujeitos acertaram acertadamente ao item. Caso contrário, sendo $ID_i = 0$ indica que o item não foi respondido de maneira correta por nenhuma das pessoas.

A respeito deste índice, ERTHAL (2010, p.70) pontua que:

O objetivo de conhecer o ID é escolher a amostra de itens que irá compor o teste: uma amostra homogênea, onde todos os itens têm o mesmo grau de dificuldade, ou uma amostra heterogênea, em que há um aumento progressivo no grau de dificuldade dos itens, sempre prevalecendo os de dificuldade média.

Ressalta-se que no âmbito da Psicometria Clássica, a ideia de dificuldade de um item tem coerência apenas no contexto de testes de aptidão, onde há respostas certas e erradas, não havendo, portanto, ambiente para tal no contexto de testes cuja resposta depende da preferência do sujeito e não da sua capacidade como, por exemplo, testes de personalidade (PASQUALI, 2009a).

¹ São itens que não se caracterizam apenas pela presença do fator certo ou errado, mas por todo um conjunto ordenado de respostas

4.2 ÍNDICE DE DISCRIMINAÇÃO

O Índice de Discriminação (IDS) tem por objetivo diferenciar os grupos de alta e baixa pontuação (ou escore). Para [Andrade \(1983, p.54\)](#) este índice é definido como “a diferença entre as médias do item calculadas no grupo superior e no grupo inferior”, sendo o grupo superior formado por 27% dos indivíduos com os escores mais altos e 27% formando o grupo inferior, composto pelos indivíduos com os escores totais mais baixos. De acordo com [ERTHAL \(2010\)](#) este índice é o indicador de consistência interna, sendo assim, “para um teste discriminar bem o grupo de participantes testados na variável medida, é necessário que apresente um bom grau de consistência interna entre os itens e a totalidade do teste ” ([ERTHAL, 2010, p.73](#)).

O IDS está no intervalo $[-1, 1]$ e corresponde a diferença entre o “índice de dificuldade dos indivíduos que obtiveram uma pontuação elevada no escore total do teste e o índice de dificuldade dos indivíduos que obtiveram uma pontuação baixa no escore total do teste” ([SARTES; SOUZA-FORMIGONI, 2013, p.23](#)). Além disso, quanto maior o valor do índice, mais discriminativo ele será.

4.3 FIDEDIGNIDADE E VALIDADE DE UM TESTE

Para garantir a cientificidade e legitimidade dos instrumentos de medidas, é necessário a empregabilidade de certos recursos que assegurem tais condições. Desta forma, tais instrumentos devem apresentar certas características que garantem a confiabilidade em seus dados. Para o caso de instrumentos físicos faz-se necessário sua calibração, o mesmo pode ser pensado para o caso da Psicometria, onde o ajuste dos testes, se dão por meio validade e fidedignidade.

De acordo com [Pasquali \(2009a, p.192\)](#) “a fidedignidade ou a precisão de um teste diz respeito à característica que ele deve possuir, a saber, medir sem erros”, isto é, o mesmo teste, quando aplicado para a um grupo de sujeitos em diferentes situações, ou testes similares, medindo os examinados na mesma ocasião, produz escores (resultados) idênticos. Destaca-se ocorrerá sempre erro de medida, isto é, variações que não representam as diferenças reais. “Portanto, uma prática corrente é estimar a precisão de um determinado teste para conseguir estabelecer uma expectativa de quão errônea poderá ser a medida” ([PRIMI, 2012, p.300](#)). Bem como pontua [Erthal \(2009\)](#) a fidedignidade é composta por três aspectos fundamentais: precisão, estabilidade e consistência interna ou homogeneidade. A precisão visa reduzir ao máximo os erros de mensuração em uma medição, a estabilidade implica em reproduzir diferentes fenômenos, acarretando pouco erro de mensuração em momentos diferentes e a consistência interna resulta em que todos os itens de um teste meçam os mesmos pontos.

Para medir a precisão de um teste, temos as seguintes formas:

Teste-reteste: é a aplicação de uma mesma medida em duas situações distintas. Por exemplo, se um indivíduo é examinado em um teste e posteriormente ele realiza novamente o

mesmo teste, é desejável que os escores (resultados) sejam similares. A aplicação deste método requer que o aspecto a ser medido permaneça o mesmo nos dois momentos do teste e qualquer alteração no resultado decorre de erros aleatórios. (KOTTNER et al., 2011)

Divisão das Metades: Particiona-se um único teste em duas metades equivalentes. Para a divisão do teste, é necessário o conhecimento a cerca da dificuldade dos itens e o “peso” atribuído, assim as partes terão grandes semelhanças. Este método não é aplicável em testes longos e que exigem um alto grau de atenção, visto que a fadiga pode acarretar perda de notas da segunda metade (ERTHAL, 2009).

Formas Paralelas: A fidedignidade do teste é obtida por meio dos escores do mesmo examinado em duas formas paralelas do mesmo teste, “a correlação com estes dois escores constitui o coeficiente de precisão. Vale ressaltar que, nas duas formas paralelas, os testes devem conter o mesmo número de itens, expressos da mesma forma e abrangendo o mesmo conteúdo” (OTTATI, 2003, p.2).

A validade de um teste pode ser compreendida como sendo a capacidade de um teste de mensurar o que foi proposto. Para Urbina (2009, p.107) “refere-se àquilo que um teste mede e a quão bem ele faz isso”. Diz-se que um instrumento é dito válido quando as diferenças entre os resultados obtidas por meio do instrumento avaliativo representam as diferenças entre os sujeitos ou entre o mesmo sujeito em momentos distintos (ERTHAL, 2009).

De modo geral, há três tipos principais de validade, são elas:

Validade de conteúdo: “é o exame sistemático do conteúdo do teste com o objetivo de verificar se este realmente constitui uma amostra representativa do comportamento que deseja mensurar” (ERTHAL, 2009, p.112). De modo geral, este tipo de validade preocupa-se em demonstrar se uma medida representa bem o construto através de avaliação de um especialista. Por exemplo, um teste desenvolvido ao ser validado, passa por uma avaliação de especialistas para a verificação de coerência do critério que deseja-se medir.

Validade de critério: visa identificar se uma medida tem alguma associação a algum critério externo. Esta categoria de validade “consiste no grau de eficácia que ele tem em predizer um desempenho específico de um sujeito” (PASQUALI, 2009b, p.996)

Validade de constructo: apura se os itens de um teste mostram o constructo avaliado, constitui-se como uma parte fundamental em um teste. Segundo Vianna (1983, p.35) este tipo de validade “possibilita determinar se o teste é a melhor medida de algo diferente do que foi pretendido medir ou, então, se fornece elementos que não possuem efetiva consequência no processo de avaliação”.

5 RESULTADOS 1

No campo educacional, um instrumento metodológico fundamental, são os relatórios pedagógicos, cuja finalidade é apresentar dados e informações referentes às avaliações. As informações obtidas por meio das 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”(SOUSA; FERREIRA, 2019b, p.16). Visto que este trabalho tem por objetivo principal interpretar a análise de itens de matemática em Relatório Pedagógico de avaliação educacional, selecionamos o que tem maior impacto nacional, que é o último Relatório do SAEB, do ano de 2017.

5.1 OS RELATÓRIOS DO SAEB

O relatório SAEB tem o objetivo de compartilhar com gestores públicos e educacionais, bem como a toda população brasileira, os resultados da aplicação das provas. Além disso, um dos propósitos deste documento, é o de ceder informações que contribuem “para o constante aprimoramento da educação básica ofertada aos estudantes brasileiros” (INEP, 2017, p.15).

Destaca-se que o relatório do SAEB, do ano de 2017, em sua composição, estrutura-se em três capítulos. O primeiro capítulo tem como título “ O sistema de avaliação da educação básica (SAEB)” e traz uma breve contextualização acerca da evolução do Saeb, as normas da edição de 2017 e os conceitos que são abordados no documento. No segundo capítulo “Metodologia do Saeb” são apresentados os tópicos da metodologia utilizada no Saeb. Já no terceiro capítulo “Os resultados do SAEB 2017” são demonstrados os resultados de desempenho atingidos na avaliação.

Um documento de grande importância para o processo de elaboração dos testes do Saeb, são as matrizes de referência.¹ Para a disciplina de matemática, o INEP estrutura os conteúdos em quatro grandes eixos, são eles: Geometria, Grandezas e Medidas, Números e Álgebra e Tratamento da informação.

Os conteúdos associados a competências e habilidades desejáveis para cada série e para cada disciplina são subdivididos em partes menores, os descritores, cada uma especificando o que os itens das provas devem medir. Os descritores, por sua vez, traduzem uma associação entre os conteúdos curriculares e as operações mentais desenvolvidas pelos alunos. Os descritores, portanto, especificam o que cada habilidade implica e são utilizados como base para a construção dos itens de diferentes disciplinas. (INEP, 2020)

Ressalta-se que, para o final do primeiro ciclo de aprendizagem, (5º ano do Ensino Fundamental) há 28 descritores. Para o final do segundo ciclo, ou seja, o 9º ano do Ensino

¹ Para mais informação, acesse: <<https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb/matrizes-e-escalas>>

Fundamental, existem 37 descritos, e para a 3ª série do Ensino Médio, estão reservados 35 descritores.

Pontua-se que parte das avaliações externas nacionais utilizam-se de metodologias estatísticas e psicométricas como, por exemplo, Teoria Clássica de Testes. Segundo o relatório do INEP (2017, p.66), por meio da TCT, são calculados:

- i) os percentuais de escolha para cada possibilidade de resposta do item (inclusive respostas em branco e inválidas), um índice que reflete o quanto o item discrimina os estudantes de maior e menor desempenho; ii) o percentual de acerto dos grupos de desempenho superior e inferior; e iii) o coeficiente bisserial de cada possibilidade de resposta (inclusive respostas em branco e inválidas).

O coeficiente bisserial (3.3), para o relatório do SAEB, é utilizado como uma medida que relaciona o desempenho do aluno no item e seu desempenho no teste. Sendo assim, quando esta medida assume um valor positivo, no item, aponta que, quanto maior a performance de um certo grupo de alunos, maior a porcentagem de acerto. Para o caso de um valor negativo dessa medida, representa que os estudantes com alto desempenho não assinalaram a alternativa correta e sim outras alternativas. No que se refere ao valor do coeficiente bisserial de cada alternativa, essa medida aponta a busca em cada uma delas. Sendo assim, se o coeficiente for positivo em uma determinada alternativa, quer dizer que os alunos com alto desempenho buscaram assinalá-la. Assim, “espera-se que o gabarito tenha coeficiente bisserial positivo e os distratores tenham coeficiente bisserial negativo” (INEP, 2017, p.57).

A Figura 5.1 exemplifica uma análise por meio da TCT, dos coeficientes estatísticos clássicos do item.

Figura 5.1 – Coeficientes estatísticos clássicos de um determinado item.

ÍNDICES						PROPORÇÕES DE RESPOSTA						COEFICIENTES BISSEERIAIS					
GAB	DIFI	DISCR	ABAI	ACIM	BIS	A	B	C	D	“	”	A	B	C	D	“	”
A	0,72	0,56	0,39	0,95	0,66	0,72	0,09	0,09	0,08	0,01	0,02	0,66	-0,40	-0,43	-0,37	-0,72	-0,53

Fonte: SAEB (2017)

GAB – Indica o gabarito do item (no caso da tabela, o gabarito é dado pela alternativa A).

Índice de Dificuldade (DIFI) – Corresponde ao percentual de respostas corretas do item. Quanto maior este índice, mais fácil o item.

Índice de Discriminação (DISCR) – É a diferença entre a proporção de acertos dos estudantes com maior escore (ACIM – 27% dos respondentes com pontuações mais altas) e a proporção de acertos daqueles com menor escore (ABAI – 27% dos respondentes com pontuações mais baixas).

Coeficiente bisserial (BIS) – Medida de correlação entre o desempenho no teste e o desempenho no item.

Proporções de resposta – Indicam os percentuais de escolha por opção de resposta. As informações relativas às respostas em branco são designadas com aspas -""- enquanto, as relativas às respostas inválidas são designadas com ponto entre aspas -".-.

Coeficientes Bisseriais – Coeficientes das correlações bisseriais entre cada opção de resposta e o escore no teste. Esses coeficientes devem ser positivos para a opção correta (gabarito) e, desejavelmente, negativos para as demais (distratores). As informações relativas às respostas em branco são designadas com aspas -""-, enquanto as relativas às respostas inválidas são designadas com ponto entre aspas -".-.

5.2 INTERPRETAÇÃO DA ANÁLISE DE ITENS DE MATEMÁTICA NO RELATÓRIO PEDAGÓGICO DO SAEB - 2017

Com o objetivo de exemplificar a análise de itens presente nos relatórios do SAEB, esta seção divide-se em em duas etapas. A primeira será reservada para a apresentação do item e a compreensão das informações gerais, bem como, a série onde o item foi aplicado, dificuldade do item, descritor, gabarito e dificuldade escala. A segunda parte contará com a interpretação da parte estatística de cada item.

5.3 INTERPRETAÇÃO GERAL

O primeiro item a ser interpretado, segundo consta no relatório do Saeb, é o de número 22227 e esta apresentado na Figura 5.2.

Figura 5.2 – Item 22227 - Matemática - 5º ano do Ensino Fundamental - SAEB 2017

A imagem representa uma estante com filmes da coleção de Roberta:

Roberta está em frente da estante e escolhe o terceiro filme, da esquerda para direita, na segunda prateleira.

Qual foi o filme escolhido?

(A) A Era do Gelo 1
 (B) A Professora Maluquinha
 (C) A Bela e a Fera
 (D) O Rei Leão 1

Fonte: SAEB (2017)

Tabela 5.1 – Informações gerais do item - 22227

Área de Conhecimento	Matemática – 5º ano do Ensino Fundamental
Dificuldade	Média
Descritor	D1
Gabarito	B

Tabela 5.2 – Fonte: Adaptada pela autora

Para que o aluno consiga responder o item de maneira correta é importante que ele considere o referencial de que a personagem está localizada em frente à estante e também que considere o referencial da segunda prateleira. Além disso, o aluno deve conhecer as direções direita e esquerda, e também o conceito de ordem (primeiro, segundo, terceiro, etc). Bem como apresentado na Tabela 5.2, este item está identificado pelo descritor D1, o conteúdo e a habilidade que estão sendo avaliados, bem como consta Matriz de Referência de Matemática do SAEB: Temas e seus descritores - 4ª série/5º ano do ensino fundamental ² é o de “identificar a localização/movimentação de objeto em mapas, croquis e outras representações gráficas” (INEP, 2017, p.28).

O segundo item a ser interpretado, de acordo com a classificação do Relatório do Saeb é o de número 34762 e está representado na Figura 5.3

Figura 5.3 – Item 34762 - Matemática - 9º ano do Ensino Fundamental - SAEB 2017

<p>Maria comprou um fogão por 240 reais e uma mesa por 180 reais, pagando 150 reais de entrada e o restante em 3 vezes sem juros.</p> <p>Qual é o valor de cada prestação?</p> <p>(A) 90 reais.</p> <p>(B) 130 reais.</p> <p>(C) 140 reais.</p> <p>(D) 190 reais.</p>

Fonte: SAEB (2017)

² Para uma melhor visualização, acesse <<https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb/matrizes-e-escalas>>

Tabela 5.3 – Informações gerais do item - 34762

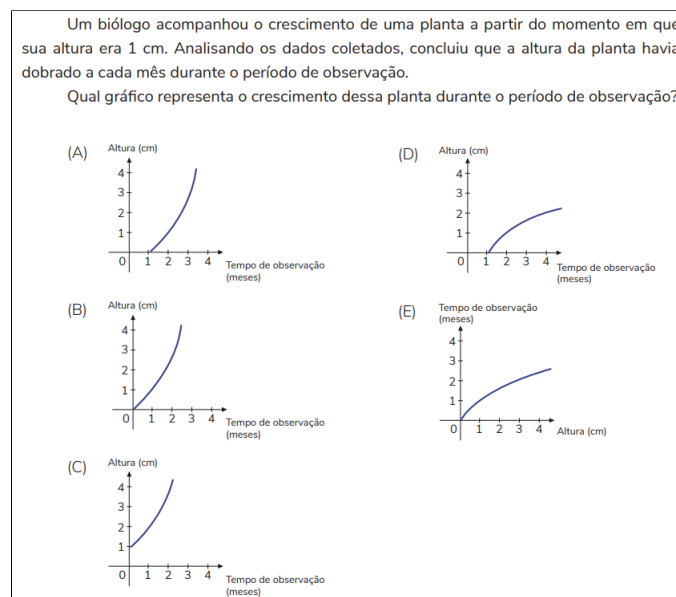
Área de Conhecimento	Matemática – 9º ano
Dificuldade	Fácil
Descritor	D19
Gabarito	A

Tabela 5.4 – Fonte: Adaptada pela autora

“Neste item, a habilidade que está sendo avaliada é a de resolver problema envolvendo números naturais e diferentes significados das operações, nesse caso, da adição, da subtração e da divisão” (INEP, 2017, p.78). Uma forma correta de resolução é somar o valor dos produtos, ou seja, $240 + 180 = 420$ e posteriormente efetuar a subtração do valor de entrada do valor total dos produtos, isto é, $420 - 150 = 270$ e por fim, realizar a divisão do resultado obtido dessa subtração pela quantidade de prestações, $270 \div 3 = 90$. O descritor neste item é representado por D19, e segundo consta na Matriz de Referência de Matemática do SAEB: Temas e seus descritores - 9ºano do ensino fundamental, avalia-se a “habilidade de Resolver problema com números naturais, envolvendo diferentes significados das operações (adição, subtração, multiplicação, divisão, potenciação)”(INEP, 2020, p.16)

O último item a ser interpretado, segundo consta no Relatório do Saeb, é o de código 29036 e está disposto na Figura 5.5.

Figura 5.4 – Código 29036 - Matemática - 3º série do Ensino Médio - SAEB 2017



Fonte: SAEB (2017)

Tabela 5.5 – Informações gerais do item - 34762

Área de Conhecimento	Matemática – 3º série do ensino médio
Dificuldade	Médio
Descritor	D21
Gabarito	C

Fonte: adaptada pela autora com base no Relatório do SAEB - 2017

“Nesse item, a habilidade avaliada é a de identificar o gráfico que representa uma situação descrita em um texto”(INEP, 2017, p.81). Para conseguir resolver corretamente, o aluno precisa notar que o eixo vertical contém a altura da planta e o eixo horizontal, o tempo de observação. Desta forma, o gráfico que possui altura de 1 cm no eixo vertical e que aponta que a altura da planta dobra a cada mês, é dado pela alternativa C. Este item tem como representação o descritor D21, que segundo consta na Matriz de Referência de Matemática do SAEB: Temas e seus descritores - 3ª série do Ensino Médio, avalia: “identificar o gráfico que representa uma situação descrita em um texto” (INEP, 2020, p.18).

Esta segunda parte da interpretação da análise dos três itens apresentados acima, visa esclarecer as estatísticas aplicadas a cada um dos itens. A Tabela 5.6 traz em sua estrutura os dados dos respectivos índices, proporções de resposta e coeficientes bisseriais, dos itens aplicados ao Ensino Fundamental. Já a Tabela 5.9 exibe as estatística do item (código) 29036, aplicado a 3ª série do Ensino Médio.

Tabela 5.6 – Coeficientes estatísticos clássicos dos itens Saeb 2017 – Matemática – Ensino Fundamental I e II

Item	TCT																		
	ÍNDICES						PROPORÇÃO DE RESPOSTA						COEFICIENTES BISSERIAIS						
	GAB	DIFI	DISCR	ABAI	ACIM	BIS	A	B	C	D	E	“”	“.”	A	B	B	D	“ ”	“.”
22227	C	0,45	0,50	0,21	0,71	0,52	0,14	0,45	0,16	0,23	0,00	0,00	0,02	-0,37	0,52	-0,26	-0,07	-0,58	-0,66
34762	A	0,71	0,50	0,44	0,94	0,59	0,71	0,13	0,10	0,05	0,00	0,00	0,00	0,59	-0,45	-0,35	-0,37	0,00	-0,49

Tabela 5.7 – Tabela adaptada pela autora com base no Relatório do SAEB - 2017

O item 22227 tem como gabarito a alternativa B. O índice de dificuldade (DIFI) obteve como resultado 0,45, dando assim um grau de dificuldade médio para o item, bem como consta na Tabela 5.2. Além disso, por (4.2), sabe-se que $0 \leq DIFI \leq 1$, logo 0,45 está mais próximo ao meio do intervalo. De modo geral, o percentual de acertos foi de 45%, o que indica que um número médio de estudantes conseguiu identificar a localização/movimentação do objeto, em uma representação gráfica. Com relação ao Índice de Discriminação (DISCR), obteve-se como resultado 0,50, o que indica que o item conseguiu separar de maneira satisfatória os grupos com melhor e pior desempenho, pois, bem como apresentado em 4.2, este índice varia de -1 a $+1$ e quanto maior o valor de DISC, mais discriminativo ele será. Os índice ABAI, indica os percentuais de acerto no grupo de pior desempenho. Para este item o percentual é de 21%. Já o

índice ACIM, indica os percentuais de acerto no grupo de melhor desempenho. Neste índice o percentual foi de 71%. O coeficiente bisserial (BIS) assumiu valor de 0,52, o que indica que este item é considerado bom. Nota-se pelos valores dos coeficientes bisseriais de cada uma das alternativas, verifica-se valores negativos. Porém, no caso do distrator D, o resultado obtido foi de $-0,07$, um valor relativamente próximo a zero, indicando que alunos com baixa proficiência assinalaram esta alternativa. Com relação às respostas em branco, obteve-se $'' = -0,58$. Já o resultado para as respostas inválidas é dado por $.'' = -0,66$. Com relação a distribuição de respostas, tem-se que o gabarito do item, dado pela letra B, computou um percentual de 45%. A alternativa A atraiu 14% dos estudantes, os quais possivelmente consideraram a segunda prateleira como sendo a segunda coluna e além disso, podem ter reconhecido o terceiro filme na orientação de cima para baixo. A alternativa C atraiu 16%, os quais possivelmente reconheceram o terceiro filme da esquerda para a direita na primeira prateleira. O item D atraiu 23% e chama a atenção pelo fato do alto percentual e também por conta de alunos de média e alta proficiência estarem neste grupo.

Para o item 34762 a resposta correta é dada pela alternativa A. O índice de dificuldade (DIFI) aponta um alto percentual de acertos, isto é, 71%, indicando que os alunos possuem a habilidade da resolução de problema envolvendo os números naturais e diferentes significados das operações. Já o índice de discriminação (DISC) obteve como resultado o valor de 0,50, apontando que o item conseguiu separar de maneira adequada os estudantes com melhor e pior desempenho, visto que quanto maior o valor de DISC, mais discriminativo ele será. Com relação aos percentuais de acerto no grupo de pior desempenho (ABAI), o percentual foi de 44%. Já os percentuais de acerto no grupo de melhor desempenho (ACIM) o percentual obtido no item foi de 94%. A correlação bisserial (BIS) assumiu valor de 0,59, indicando que o gabarito do item não "atraiu" estudantes com baixa proficiência. Isso pode ser verificado pelos valores negativos dos coeficientes bisseriais dos distratores, evidenciando que eles atraíram apenas alunos com baixo rendimento. Com relação à distribuição de resposta, obteve-se (A)71%, um alto índice de respostas corretas. A alternativa B atraiu 13% dos estudantes, os quais possivelmente somaram o preço do fogão ao valor da entrada ($240+150 = 390$) e posteriormente efetuaram a divisão do resultado pela quantidade total de prestação ($390/3 = 130$). A alternativa C atraiu 10% dos estudantes, os quais possivelmente adicionaram os preços dos produtos ($240 + 180 = 420$) e, posteriormente, dividiram o resultado pela quantidade total de prestações ($420 \div 3 = 140$). Já a alternativa D atraiu 5% dos estudantes, os quais possivelmente adicionaram todos os valores ($240 + 180 + 150 = 570$) e assim dividiram o resultado pela quantidade de prestações ($570 \div 3 = 190$). Destaca-se que neste item não ocorreram respostas em branco, visto que $'' = 0,00$. Já o resultado para as respostas inválidas é dado por $.'' = -0,49$.

A Tabela 5.9 exibe os resultados referente ao item aplicado para a 3ª série do Ensino Médio, no qual é identificado pelo código 29036. Diferente dos itens do Ensino Fundamental que em sua estrutura possuem quatro alternativas, os itens para o Ensino Médio possuem cinco opções de respostas.

Tabela 5.8 – Coeficientes estatísticos clássicos dos itens Saeb – Matemática – 2017

TCT																				
Item	ÍNDICES						PROPORÇÃO DE RESPOSTA						COEFICIENTES BISSERIAIS							
	GAB	DIFI	DISCR	ABAI	ACIM	BIS	A	B	C	D	E	""	“.”	A	B	C	D	E	“.”	“.”
29036	C	0,49	0,57	0,22	0,79	0,57	0,20	0,14	0,49	0,08	0,08	0,00	0,00	-0,29	-0,27	0,57	-0,28	-0,27	-0,59	0,00

Tabela 5.9 – Tabela adaptada pela autora com base no Relatório do SAEB - 2017

Bem como consta na Tabela 5.9, o gabarito para este item é dado pela alternativa C. O valor do índice de Dificuldade (DIFI), dado por 0,49, indica um grau de dificuldade médio, ou seja, 49% dos estudantes acertaram a resposta correta. O Índice de Discriminação (DISCR), obteve como resultado o valor de 0,57, sendo assim classificado como satisfatório, indicando que o item conseguiu distinguir os estudantes que obtiveram melhor rendimento. Com relação ao índice de que aponta os percentuais de acerto no grupo de pior desempenho (ABAI), o resultado obtido no item foi de 22%. O índice que aponta os percentuais de acerto no grupo de melhor desempenho (ACIM), o resultado obtido foi de 79%. O coeficiente bisserial (BIS) obteve como resultado o valor de 0,57 indicando que o item não atraiu estudantes com baixa proficiência. Esse fator também pode ser verificado pelos valores negativos dos coeficientes bisseriais dos distratores, evidenciando que eles atraíram apenas alunos de baixo rendimento. A proporção de de resposta em branco foi de -0,59. Destaca-se que neste item não houve respostas inválidas.

Com relação as proporções de respostas, nota-se que a alternativa (A) atraiu 20% dos estudantes, os quais podem ter considerado a altura de 1 cm no eixo incorreto. A alternativa B atraiu 14% dos estudantes, os quais provavelmente consideraram o acompanhamento do crescimento da planta com início na altura zero. A alternativa D atraiu 8% dos estudantes, os quais podem possivelmente considerado a altura de 1 cm no eixo incorreto e que a altura da planta não aumenta de forma exponencial. Já a alternativa E também atraiu 8% dos estudantes, uma justificativa para a escolha desta alternativa pode ser o fato de que os estudantes consideraram o acompanhamento do crescimento da planta a partir da altura zero.

6 RESULTADO 2

Uma das exigências da sociedade atual é que a educação prepare o aluno para lidar com situações distintas a cada dia. Visto que as tecnologias se fazem cada vez mais integradas ao nosso cotidiano, é esperado que a escola tenha que se atualizar, se almeja sobreviver como instituição educacional. Desta forma, é fundamental “que o professor se aproprie de gama de saberes advindos com a presença das tecnologias digitais da informação e da comunicação para que estes possam ser sistematizadas em sua prática pedagógica” (SOUSA et al., 2011, p.20).

Com o objetivo de apoiar professor e gestores nos cálculos envolvidos na análise de itens em avaliações educacionais, foi desenvolvida uma planilha eletrônica que calcula alguns indicadores da Teoria Clássica de Testes, partindo de um banco de respostas binárias da prova de Matemática do ENEM edição 2019. A Seção 6.1 aborda detalhes dessa planilha.

6.1 A PLANILHA

O *Google Sheets* ou no português “Planilhas Google” é uma ferramenta de planilhas *online* inclusa no pacote gratuito de edição de documentos oferecido pelo *Google*¹, permitindo que o usuário crie, formate arquivos e colabore com outras pessoas em tempo real. Este serviço está disponível como aplicativo da *WEB*, aplicativo móvel para celulares e também como aplicativo de *desktop* no *Chrome OS* do *Google*.

Para a implementação das planilhas interativas, o orientador forneceu, a partir de suas pesquisas, uma amostra de 200 examinados do banco de respostas da prova de Matemática do ENEM 2019, com o filtro "cidade de São Carlos AND Prova Amarela", resultando em uma matriz de dimensão 200x45. Em seguida foi criada uma planilha contendo as respostas no formato binário, sendo 1 = resposta correta e 0 = resposta incorreta, bem como exemplificado na Figura 6.1.

¹ Empresa multinacional de serviços online e software, que desenvolve e armazena inúmeros serviços e produtos baseados na internet. Para mais informação, acesse: <https://about.google/?utm_source=google-BRutm_medium=referralutm_campaign=hp-footerfg=1>

Figura 6.1 – Recorte da planilha interativa no *Google Sheets*

Sujeito	Item 136	Item 137	Item 138	Item 139	Item 140
1	1	0	0	0	0
2	1	0	0	1	1
3	0	0	0	0	0
4	0	1	0	0	0
5	1	0	1	1	0
6	0	0	0	0	0
7	1	0	0	1	0
8	0	1	0	0	0
9	1	1	0	1	0
10	0	0	0	0	1

Fonte: Planilha elaborada pela autora

Uma vez que o professor inseriu as respostas dos 200 examinados, a planilha interativa passa a calcular alguns indicadores e índices da TCT, conforme recorte da tabela mostrada na Figura 6.2.

Figura 6.2 – Cálculo das estatísticas para interpretação dos itens no *Google Sheets*

	A	B	C	D	E
1		Item 136	Item 137	Item 138	Item 139
2	Total de acertos	65	49	45	119
3	Total de erros	135	151	155	81
4	Total de respostas	200	200	200	200
5	Índice de Facilidade (IF)	0,33	0,25	0,23	0,60
6	Média de Erros no item (1-IF)	0,68	0,76	0,78	0,41
7	Classificação	Difícil	Muito difícil	Muito difícil	Fácil
8	Quantidade de acertos grupo superior	34	22	25	46
9	Quantidade de acertos grupo inferior	7	5	7	17
10	IF Superior	0,63	0,41	0,46	0,85
11	IF Inferior	0,13	0,09	0,13	0,31
12	ACIM	0,63	0,41	0,46	0,85
13	ABAI	0,13	0,09	0,13	0,31
14	Índice de Discriminação (DISCR)	0,50	0,31	0,33	0,54
15	Média de acertos total no item (Xa)	13,15	13,00	13,38	13,10
16	Média de pessoas que erraram o teste (Xm)	11,49	12,46	12,04	10,17
17	Desvio padrão (S)	6,43	6,43	6,43	6,43
18	IF*(1-IF)	0,22	0,18	0,17	0,24
19	IF/(1-IF)	0,48	0,32	0,29	1,47
20	Raiz(IF/(1-IF))	0,69	0,57	0,54	1,21
21	Rbiserial	0,18	0,05	0,11	0,55

Fonte: Elaborada pela autora

A interpretação da coluna B da planilha interativa da Figura 6.2 é da seguinte forma:

Total de acertos: total de acertos do item.

Total de erros: total de erros do item.

Total de respostas: quantidade total de respostas de acordo com o número de indivíduos que realizaram a prova.

Índice de facilidade (IF): corresponde ao percentual de respostas corretas no item. (4.2).

Média de Erros no item (1-IF): Calcula a média de erros em cada item, ou seja - Equação $(1 - IF)$.

Classificação: classificação do grau de dificuldade do item com base no resultado do (4.2) no intervalo de $[0, 1]$.

Quantidade de acertos grupo superior: quantidade de acertos no grupo que contém os indivíduos enquadrados nos 27% superiores da distribuição.

Quantidade de acertos grupo inferior: quantidade de acertos no grupo que contém os indivíduos enquadrados nos 27% inferiores da distribuição.

IF Superior: índice de facilidade no grupo superior. É obtido por meio da divisão entre a quantidade de acertos no grupo de melhor desempenho pela quantidade de indivíduos que pertencem ao grupo de melhor desempenho.

IF Inferior: índice de facilidade no grupo de pior desempenho. É obtido por meio da divisão entre a quantidade de acertos no grupo de pior desempenho pela quantidade de indivíduos que pertencem ao grupo de pior desempenho.

ACIM: índice que indica o percentual de acerto no grupo de melhor desempenho.

ABAI: índice que indica o percentual de acerto no grupo de pior desempenho.

Índice de Discriminação (ID): diferencia os grupos de alta e baixa pontuação. É calculado por meio da diferença entre o IF do grupo de melhor desempenho com o IF do grupo de pior desempenho.

Média de acertos total no item (Xa): é obtido por meio da média das notas dos sujeitos que acertaram o item.

Média de pessoas que erraram o teste (Xm): é obtido por meio da média das notas dos sujeitos que erraram o item.

Desvio padrão (S): desvio padrão das notas dos indivíduos.

IF * (1-IF): produto do índice de facilidade no item pela quantidade de erros no item.

IF/(1-IF): quociente do índice de facilidade no item pela quantidade de erros no item.

Raiz (IF/(1-IF)): raiz do quociente do índice de facilidade no item pela quantidade de erros no item.

Rbisserial: resultado do coeficiente bisserial por pontos em cada item.

Para apoiar o cálculo da correlação ponto-bisserial, foi desenvolvida uma planilha interativa, conforme mostra a Figura 6.3. Para a implementação dessa planilha interativa, o orientador forneceu, a partir de suas pesquisas, as opções de resposta de cada um dos 200 examinados da prova de Matemática do ENEM 2019, com o filtro "cidade de São Carlos AND Prova Amarela", resultando em uma matriz de dimensão 200x45.

Figura 6.3 – Análise do vetor-resposta no *Google Sheets*

Gabartio	B	E	D	E	E
Sujeito	Item 136	Item 137	Item 138	Item 139	Item 140
1	B	A	A	C	B
2	B	B	C	E	E
3	E	B	A	B	C
4	E	E	A	B	D
5	B	D	D	E	B
6	D	B	C	D	A
7	B	A	A	E	C
8	C	E	C	D	C
9	B	E	B	E	C
10	A	D	C	A	E

Fonte: Elaborada pela autora

Com base na Tabela 5.1 foram calculadas as proporções de respostas e os coeficientes bisserias para cada alternativa do item. A Figura 6.4 traz o recorte dos cálculos realizados.

Figura 6.4 – Frequência na escolha das alternativas do item *Google Sheets*

	A	B	C	D	E
1	Escolha	Item 136	Item 137	Item 138	Item 139
2	A	17	63	46	14
3	B	65	45	36	21
4	C	41	24	50	13
5	D	51	18	45	33
6	E	25	49	22	119
7	.	1	1	1	0
8	*				
9	% Escolha	Item 136	Item 137	Item 138	Item 139
10	A	8,50%	31,50%	23,00%	7,00%
11	B	32,50%	22,50%	18,00%	10,50%
12	C	20,50%	12,00%	25,00%	6,50%
13	D	25,50%	9,00%	22,50%	16,50%
14	E	12,50%	24,50%	11,00%	59,50%
15	.	0,50%	0,50%	0,50%	0,00%
16					
17	Índice de escolha	Item 136	Item 137	Item 138	Item 139
18	A	0,09	0,32	0,23	0,07
19	B	0,33	0,23	0,18	0,11
20	C	0,21	0,12	0,25	0,07
21	D	0,26	0,09	0,23	0,17
22	E	0,13	0,25	0,11	0,60
23	.	0,01	0,01	0,01	0,00
24					

Fonte: Elaborada pela autora

A interpretação da coluna B da planilha iterativa da Figura 6.4 é da seguinte forma:

Escolha: quantidade de escolhas na alternativa do item.

% Escolha: percentual de escolhas na alternativa do item.

Índice de escolha: é obtido por meio da divisão da escolha pela quantidade de sujeitos que realizaram o teste.

Em seguida, calculou-se a média de acertos em cada alternativa do item, bem como consta na Figura 6.5:

Figura 6.5 – Cálculo da média de acertos em cada alternativa no item *Google Sheets*

25	Med acertos A	14,94	14,71	12,69	15,36
26	Med de erros A	14	11,60	10,33	10,21
27					
28	Med acertos B	13,89	13,38	15,58	12,19
29	Med de erros B	17,31	12,87	11,83	11,90
30					
31	Med acertos C	11,95	13,71	12,42	11,50
32	Med de erros C	11,46	12,96	13,82	9,54
33					
34	Med acertos D	13,63	12,50	13,18	14,36
35	Med de erros D	10,59	13,44	18,00	9,30
36					
37	Med acertos E	12,36	11,84	13,73	13,28
38	Med de erros E	11,72	16,20	11,59	15,56

Fonte: Elaborada pela autora

Para a obtenção do coeficiente bisserial de cada alternativa no item, foram utilizados os mesmos índices calculados na planilha de resposta binária. A Figura 6.6 exemplifica este caso:

Figura 6.6 – Cálculo do coeficiente bisserial para cada alternativa *Google Sheets*

40	Desvio (P)	6,38	6,38	6,38	6,38
41	IF	0,33	0,25	0,23	0,60
42	Erro no Item (q)	0,68	0,76	0,78	0,41
43	IF*q	0,22	0,18	0,17	0,24
44	IF/q	0,48	0,32	0,29	1,47
45	Raiz(IF/q)	0,69	0,57	0,54	1,21
46					
47	Rbiss A	0,10	0,28	0,20	0,98
48	Rbiss B	-0,37	0,05	0,32	0,05
49	Rbiss C	0,05	0,07	-0,12	0,37
50	Rbiss D	0,33	-0,08	-0,41	0,96
51	Rbiss E	0,07	-0,39	0,18	-0,43

Fonte: Elaborada pela autora

6.2 ESTUDO DE CASO

O critério para a escolha dos itens para esta monografia foi: item mais fácil e mais difícil.

O primeiro item a ser analisado, de acordo com a classificação do caderno amarelo do ENEM do ano de 2019, é o de número 148 e está representado na Figura 6.7:

Figura 6.7 – Item 148 — Matemática — ENEM 2019 — Caderno Amarelo

Questão 148

Um professor aplica, durante os cinco dias úteis de uma semana, testes com quatro questões de múltipla escolha a cinco alunos. Os resultados foram representados na matriz.

$$\begin{bmatrix} 3 & 2 & 0 & 1 & 2 \\ 3 & 2 & 4 & 1 & 2 \\ 2 & 2 & 2 & 3 & 2 \\ 3 & 2 & 4 & 1 & 0 \\ 0 & 2 & 0 & 4 & 4 \end{bmatrix}$$

Nessa matriz os elementos das linhas de 1 a 5 representam as quantidades de questões acertadas pelos alunos Ana, Bruno, Carlos, Denis e Érica, respectivamente, enquanto que as colunas de 1 a 5 indicam os dias da semana, de segunda-feira a sexta-feira, respectivamente, em que os testes foram aplicados.

O teste que apresentou maior quantidade de acertos foi o aplicado na

A segunda-feira.
B terça-feira.
C quarta-feira.
D quinta-feira.
E sexta-feira.

Fonte: ENEM (2019)

Tabela 6.1 – Informações gerais do item — 148

Área de Conhecimento	Matemática
Dificuldade calculada	Fácil
Opção correta	A

Fonte: o trabalho

O gabarito correto do item é a letra A. Uma forma correta de resolver este item é observar que sendo os elementos das linhas de 1 a 5 as quantidades de questões acertadas por Ana, Bruno, Carlos, Denis e Érica, respectivamente, e as colunas de 1 a 5 indicando os dias da semana, de segunda a sexta-feira, então (i) quantidade de acertos na segunda-feira: $3 + 3 + 2 + 3 + 0 = 11$, (ii) quantidade de acertos na terça-feira: $2 + 2 + 2 + 2 + 2 = 10$, (iii) quantidade de acertos na quarta-feira: $0 + 4 + 2 + 4 + 0 = 10$, (iv) quantidade de acertos na quinta-feira: $1 + 1 + 3 + 1 + 4 = 10$

e (v) quantidade de acertos na sexta-feira: $2 + 2 + 2 + 0 + 4 = 10$. Desta forma, a maior quantidade de acertos foi aplicado na segunda-feira, na alternativa A.

O segundo item a ser interpretado é o de número 165, segundo o caderno amarelo do ENEM de 2019 e está representado na Figura 6.8

Figura 6.8 – Item 165 — Matemática — ENEM 2019 — Caderno Amarelo

Questão 165

O preparador físico de um time de basquete dispõe de um plantel de 20 jogadores, com média de altura igual a 1,80 m. No último treino antes da estreia em um campeonato, um dos jogadores desfalcou o time em razão de uma séria contusão, forçando o técnico a contratar outro jogador para recompor o grupo.

Se o novo jogador é 0,20 m mais baixo que o anterior, qual é a média de altura, em metro, do novo grupo?

A 1,60
 B 1,78
 C 1,79
 D 1,81
 E 1,82

Fonte: ENEM (2019)

Tabela 6.2 – Informações gerais do item - 165

Área de Conhecimento	Matemática
Dificuldade calculada	Difícil
Gabarito	C

Fonte: o trabalho

Este item aborda o conceito de média aritmética simples. Uma forma correta de resolução é analisar que em uma seleção com 20 jogadores, a média das alturas é de 1,8 m, para isso, basta fazer o seguinte cálculo: $\frac{x_1 + x_2 + \dots + x_{20}}{20} = 1,8$. Logo, a soma de todas as alturas é dada pelo produto $20 \cdot 1,8 = 36$ m. Visto que o novo jogador é 20cm mais baixo que o jogador que será substituído, a soma das alturas passará a ser 35,8 m. Assim, a nova média da altura dos jogadores será: $\frac{35,8}{20} = 1,79$ m. Logo, a alternativa correta é dada pela alternativa C.

A tabela 5.1 resume os índices, as proporções de respostas e os coeficientes bisserais dos itens analisados para este Estudo de Caso.

Tabela 6.3 – Resultados quantitativos do estudo de caso do ENEM 2019 – Matemática – Caderno amarelo

Item	TCT																		
	ÍNDICES						PROPORÇÃO DE RESPOSTA						COEFICIENTES BISSERIAIS						
	GAB	DIFI	DISCR	ABAI	ACIM	BIS	A	B	C	D	E	“.”	“.”	A	B	C	D	E	“.”
148	A	0,65	0,50	0,41	0,91	0,62	0,65	0,23	0,04	0,03	0,06	0,00	0,00	-0,33	0,55	0,40	0,71	0,96	0,00
165	C	0,30	0,59	0,06	0,65	0,14	0,32	0,33	0,30	0,02	0,04	0,00	0,00	0,38	-0,04	-0,37	-0,25	0,17	0,00

Tabela 6.4 – Adaptada pela autora

A resposta correta do item 148 é dado pela opção A. O valor DIFI tem como resultado 0,65, ou seja, é um item fácil. Com relação ao DISCR, o resultado obtido foi 0,50, o que indica que o item conseguiu discriminar bem os grupos de melhor e pior desempenho. Para este item, o índice ABAI teve como resultado 0,41, já o índice ACIM, o percentual foi de 0,91. O BIS assumiu valor 0,62, indicando que este é um bom item.

Para o item 165 o gabarito é dado pela alternativa C. O valor obtido no DIFI foi de 0,30, dando um grau de dificuldade difícil para o item. Já para o caso do DISCR o resultado obtido foi 0,59, indicando que o item conseguiu discriminar de maneira adequada os grupos com melhor e pior desempenho. O índice ACIM teve como valor 0,65 e o índice ABAI obteve 0,06. No caso do coeficiente bisserial (BIS) o valor obtido foi 0,14 indicando que este não é um bom item, pelo baixo valor obtido.

7 CONCLUSÃO

Os estudos de [Silva e Darsie \(2008\)](#) apontam “dificuldades apresentadas pelos professores de matemática na elaboração dos relatórios avaliativos de seus alunos” ([SILVA; DARSIE, 2008](#), p.16). Uma vez que há a necessidade, dos gestores e professores, em compreender as estatísticas e coeficientes quanto à análise de itens das avaliações externas.

Com o objetivo de preencher esta lacuna de estudos da TCT sob a ótica de professores de matemática, este TCC abordou aspectos teóricos da avaliação, testes educacionais e do SAEB. Visando instrumentalizar o professor, foram abordadas algumas estatísticas para análise de itens com foco em correlações e medidas de posição e de dispersão. Enquanto teoria educacional e psicométrica, a Teoria Clássica de Testes foi abordada com o propósito de facilitar a interpretação quantitativa e qualitativa dos itens.

Instanciando a aplicação e análise de itens via TCT, um Relatório do SAEB foi interpretado no que se refere a itens de matemática, mostrando a importância deste relatórios pedagógicos. Na continuidade, foram implementadas planilhas dinâmicas que calculam automaticamente determinados índices da TCT. Um estudo de caso do ENEM 2019 foi conduzido a partir de uma amostra de resposta daquela prova de Matemática, em que a planilha foi aplicada e os itens analisados individualmente.

Apesar destes relatórios serem publicados pelo INEP como se fossem de fácil interpretação por professores e/ou gestores, o processo metodológico deste trabalho nos indicou que somente especialistas em análise de itens e/ou estatística educacional são capazes de interpretá-los naturalmente. Sendo assim, este TCC foi desafiador, na medida em que a maior parte dos conhecimentos envolvidos nessa área de avaliação educacional não faz parte da matriz curricular do curso de licenciatura em matemática.

Além disso, percebemos que a leitura e interpretação de relatórios pedagógicos foi um ponto de grande dificuldade para o desenvolvimento desta monografia, visto que tais documentos não detalham as ferramentas que foram utilizadas para a análise.

REFERÊNCIAS

- ANDRADE, F. L. de. Considerações sobre o índice de discriminação de itens em testes educacionais. **Educação e Seleção**, n. 07, p. 54–58, 1983. Citado na página 25.
- ANDRADE, J. M. de; LAROS, J. A.; GOUVEIA, V. V. O uso da teoria de resposta ao item em avaliações educacionais: diretrizes para pesquisadores. **Avaliação Psicológica**, Instituto Brasileiro de Avaliação Psicológica, v. 9, n. 3, p. 421–435, 2010. Citado na página 21.
- BRASIL. INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA - INEP. **Relatório Saeb 2017**. Disponível em: <<https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb/matrizes-e-escalas>>. Acesso em: 20 out. 2021. Citado 5 vezes nas páginas 27, 28, 31, 32 e 33.
- BRASIL. INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA - INEP. **Sistema de Avaliação da Educação Básica (Saeb) - Apresentação**. Disponível em: <<https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb/matrizes-e-escalas>>. Acesso em: 20 out. 2021. Citado 3 vezes nas páginas 27, 32 e 33.
- BRASIL. INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA - INEP. **Exame Nacional do Ensino Médio (ENEM) - Apresentação**. Disponível em: <<https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem>>. Acesso em: 05 mar. 2021. Citado 2 vezes nas páginas 10 e 14.
- BRASIL. INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA - INEP. **PORTARIA Nº 10, DE 8 DE JANEIRO DE 2021**. Disponível em: <<https://www.in.gov.br/en/web/dou/-/portaria-n-10-de-8-de-janeiro-de-2021-298322305>>. Acesso em: 22 set.2021. Citado na página 16.
- BRASIL. INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA - INEP. **Sistema de Avaliação da Educação Básica (Saeb) - Apresentação**. Disponível em: <<https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb>>. Acesso em: 15 ago. 2021. Citado na página 16.
- CAMPBELL, D. T.; STANLEY, J. C. **Experimental and quasi-experimental designs for research**. [S.l.]: Ravenio Books, 2015. Citado na página 23.
- CHUIEIRE, M. S. F. Concepções sobre a avaliação escolar. **Estudos em avaliação educacional**, v. 19, n. 39, p. 49–64, 2008. Citado na página 12.
- DEVELLIS, R. F. Classical test theory. **Medical care**, JSTOR, p. S50–S59, 2006. Citado na página 24.
- ERTHAL, T. C. Manual de psicometria. ed. **Jorge Zaha Ltda. 8ª ed. Rio de Janeiro**, 2009. Citado 2 vezes nas páginas 25 e 26.
- ERTHAL, T. C. S. **Manual de psicometria**. [S.l.]: Zahar, 2010. Citado 2 vezes nas páginas 24 e 25.
- FUNDAÇÃO PARA DESENVOLVIMENTO DA EDUCAÇÃO - FDE. **Saresp permite monitorar avanços da educação básica no Estado**. Disponível em: <<https://saresp.fde.sp.gov.br/>>. Acesso em: 05 mar. 2021. Citado na página 14.

FIGUEIREDO, V. L. M. d. et al. Propriedades psicométricas dos itens do teste wisc-iii. **Psicologia em estudo**, SciELO Brasil, v. 13, p. 585–592, 2008. Citado na página 24.

FREIRE, P. **A importância do ato de ler em três artigos que se completam: Volume 22**. [S.l.]: Cortez editora, 2017. v. 22. Citado na página 4.

GRÉGOIRE, J.; LAVEAULT, D. Introdução às teorias dos testes em ciências humanas. **Portugal: Porto**, 2002. Citado na página 24.

GULLIKSEN, H. **Theory of mental tests**. [S.l.]: Routledge, 2013. Citado na página 23.

HAYDT, R. C. C. **Curso de didática geral**. [S.l.]: 1ª Edição-São Paulo: Ática, 2011. Citado 3 vezes nas páginas 12, 13 e 14.

KOTTNER, J. et al. Guidelines for reporting reliability and agreement studies (grras) were proposed. **International journal of nursing studies**, Elsevier, v. 48, n. 6, p. 661–671, 2011. Citado na página 26.

MACHADO, C. Avaliação externa e gestão escolar: reflexões sobre usos dos resultados. **Revista@mbienteeducação**, v. 5, n. 1, p. 70–82, 2017. Citado na página 13.

MARINHO-ARAÚJO, C. M.; RABELO, M. L. Avaliação educacional: a abordagem por competências. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, SciELO Brasil, v. 20, p. 443–466, 2015. Citado na página 12.

MIQUELANTE, M. A. et al. As modalidades da avaliação e as etapas da sequência didática: articulações possíveis. **Trabalhos em Linguística Aplicada**, SciELO Brasil, v. 56, p. 259–299, 2017. Citado 2 vezes nas páginas 12 e 13.

MORETTIN, P. A.; BUSSAB, W. O. **Estatística básica**. [S.l.]: Saraiva Educação SA, 2017. Citado 3 vezes nas páginas 17, 18 e 21.

OSTERLIND, S. J. **What Is Constructing Test Items?** [S.l.]: Springer, 1998. Citado na página 15.

OTTATI, F. Parâmetros psicométricos de instrumentos de interesse profissional. **Estudos e Pesquisas em Psicologia**, v. 3, n. 2, p. 22–31, 2003. Citado na página 26.

PASQUALI, L. Psicometria. **Revista da Escola de Enfermagem da USP**, FapUNIFESP (SciELO), v. 43, n. spe, p. 992–999, dez. 2009. Disponível em: <<https://doi.org/10.1590/s0080-62342009000500002>>. Citado 3 vezes nas páginas 23, 24 e 25.

PASQUALI, L. Psychometrics. **Revista da Escola de Enfermagem da USP**, SciELO Brasil, v. 43, p. 992–999, 2009. Citado na página 26.

PITON-GONÇALVES, J. **Desafios e perspectivas da implementação computacional de testes adaptativos multidimensionais para avaliações educacionais**. Tese (Doutorado) — Universidade de São Paulo, 2012. Citado 2 vezes nas páginas 13 e 14.

PRIMI, R. Psicometria: fundamentos matemáticos da teoria clássica dos testes. **Avaliação Psicológica**, Instituto Brasileiro de Avaliação Psicológica, v. 11, n. 2, p. 297–307, 2012. Citado 2 vezes nas páginas 23 e 25.

SARTES, L. M. A.; SOUZA-FORMIGONI, M. L. O. d. Avanços na psicometria: da teoria clássica dos testes à teoria de resposta ao item. **Psicologia: Reflexão e Crítica**, SciELO Brasil, v. 26, p. 241–250, 2013. Citado 3 vezes nas páginas 23, 24 e 25.

SILVA, M. J. da; DARSIE, M. M. P. Da nota ao relatório descritivo avaliativo: dificuldades dos professores de matemática. **Série-Estudos-Periódico do Programa de Pós-Graduação em Educação da UCDB**, 2008. Citado 2 vezes nas páginas 10 e 45.

SOARES, T. M. Utilização da teoria da resposta ao item na produção de indicadores sócio-econômicos. **Pesquisa operacional**, SciELO Brasil, v. 25, p. 83–112, 2005. Citado na página 20.

SOUSA, C. P. d.; FERREIRA, S. L. Avaliação de larga escala e da aprendizagem na escola: um diálogo necessário. **Psicologia da Educação**, Pontifícia Universidade Católica de São Paulo, n. 48, p. 13–23, 2019. Citado na página 14.

SOUSA, C. P. de; FERREIRA, S. L. Avaliação de larga escala e da aprendizagem na escola: um diálogo necessário. **Psicologia da Educação. Programa de Estudos Pós-Graduados em Educação: Psicologia da Educação. ISSN 2175-3520**, n. 48, 2019. Citado 2 vezes nas páginas 10 e 27.

SOUSA, R. P. d. et al. **Tecnologias digitais na educação**. [S.l.]: Eduepb, 2011. Citado na página 36.

TARAS, M. De volta ao básico: definições e processos de avaliação. **Práxis Educativa**, Universidade Estadual de Ponta Grossa, v. 5, n. 2, p. 123–130, 2010. Citado na página 13.

URBINA, S. **Fundamentos da testagem psicológica**. [S.l.]: Artmed Editora, 2009. Citado na página 26.

VIANNA, H. M. Validade de construto em testes educacionais. **Educação e seleção**, n. 08, p. 35–44, 1983. Citado na página 26.

Exceto quando indicado o contrário, a licença deste item é descrito como
Attribution-NonCommercial-NoDerivs 3.0 Brazil

