UNIVERSIDADE DE SÃO PAULO
FACULDADE DE EDUCAÇÃO

ÉRICA MARIA TOLEDO CATALANI

Teste Adaptativo Informatizado da Provinha Brasil: a construção de um instrumento de apoio para professores(as) e gestores(as) de escolas

São Paulo
2019

ÉRICA MARIA TOLEDO CATALANI

Teste Adaptativo Informatizado da Provinha Brasil: a construção de um instrumento de apoio para professores(as) e gestores(as) de escolas

Versão Corrigida

Tese apresentada como exigência para obtenção do título de Doutor em Educação junto ao Programa de Pós-Graduação em Educação da Faculdade de Educação da Universidade de São Paulo (Feusp)

Área de concentração: Estado, Sociedade e Educação

Orientador: Prof. Dr. Ocimar Munhoz Alavarse

São Paulo
2019

Catalogação da Publicação
Ficha elaborada pelo Sistema de Geração Automática a partir de dados fornecidos pelo(a) autor(a)
Bibliotecária da FE/USP: Nicolly Soares Leite - CRB-8/8204

CCat35
7t

CATALANI, Érica Maria Toledo
Teste Adaptativo Informatizado da Provinha Brasil / Érica Maria Toledo CATALANI; orientador Ocimar Munhoz ALAVARSE. -- São Paulo, 2019.
282 p.

Tese (Doutorado - Programa de Pós-Graduação Estado, Sociedade e Educação) -- Faculdade de Educação, Universidade de São Paulo, 2019.

1. Provinha Brasil. 2. Avaliação Educacional. 3. Teste Adaptativo Informatizado (TAI). 4. Proficiência em leitura. 5. Alfabetização. I. ALAVARSE, Ocimar Munhoz, orient. II. Título.

CATALANI, Érica Maria Toledo. **Teste Adaptativo Informatizado da Provinha Brasil: a construção de um instrumento de apoio para professores(as) e gestores(as) de escolas**. 2019. 282 p. Tese (Doutorado em Educação) – Faculdade de Educação, Universidade de São Paulo, São Paulo, 2019.

Aprovado em: 29/03/2019

Banca Examinadora

Prof. Dr. **Dalton Francisco de Andrade**

Instituição: **Universidade Federal de Santa Catarina**

Julgamento: _____

Prof. Dr. **Erica Castilho Rodrigues**

Instituição: **Universidade Federal de Ouro Preto**

Julgamento: _____

Prof. Dr. **Mariana Curi**

Instituição: **ICMC-USP**

Julgamento: _____

Prof. Dr. **Valéria Aparecida de Souza Siqueira**

Instituição: **Universidade Paulista**

Julgamento: _____

DEDICATÓRIA

Ao companheiro Ernani e às minhas filhas Maria Cecília e Fernanda, por terem compreendido e respeitado minha ausência e me incentivado durante todo o percurso.

AGRADECIMENTOS

Ao Prof. Dr. Ocimar Munhoz Alavarse, que, com generosidade, conhecimento e história de vida, ensinou-me o processo coletivo de pesquisa, orientando e reiteradamente contribuindo para meu desenvolvimento como pessoa, profissional e na vida acadêmica.

Aos(às) amigos(as) pesquisadores(as) do Gepave, por terem partilhado momentos de rico aprendizado na temática da medida e da avaliação educacional, em especial, aos pesquisadores: Rodrigo Travitzki, Douglas de Rizzo Meneghetti e Ailton Carlos Santos, por contribuírem, respectivamente, com a construção do algoritmo, o desenvolvimento da plataforma e a coordenação da equipe de formação em avaliação e alfabetização do Projeto TAI da Provinha Brasil – Leitura, decisivos para este trabalho.

Aos(às) professores(as) das disciplinas do curso, Dra. Sandra Zakia Lian Sousa, Dra. Maria Eugénia Ferrão, Dr. José de Oliveira Siqueira e Dr. Hugo Cogo Moreira, por promoverem momentos de reflexão e construção de conhecimentos acerca das diferentes dimensões da avaliação e da estatística aplicada à psicologia e à educação.

Ao Prof. Dr. Dalton Francisco de Andrade e à Profa. Dra. Mariana Curi, pelas preciosas e assertivas contribuições no exame de qualificação.

Às equipes gestoras das duas Diretorias Regionais de Educação e às equipes gestoras, docentes e discentes das 15 escolas que participaram da pesquisa, pela paciência com o processo e envolvimento na busca de novos horizontes para a avaliação educacional brasileira.

Aos parceiros de fomento ao projeto TAI da PB – Leitura, Núcleo Técnico de Avaliação da Secretaria Municipal de Educação e Instituto Nacional de Estudos e Pesquisas Anísio Teixeira, sem os quais esta investigação não seria levada a cabo.

RESUMO

CATALANI, Érica Maria Toledo. **Teste Adaptativo Informatizado da Provinha Brasil: a construção de um instrumento de apoio para professores(as) e gestores(as) de escolas**. 2019. 282 p. Tese (Doutorado em Educação) – Faculdade de Educação, Universidade de São Paulo, São Paulo, 2019.

Esta Tese resulta de um projeto de construção de um Teste Adaptativo Informatizado (TAI) para a versão em papel e lápis da Provinha Brasil (PB), focado na avaliação da proficiência em leitura. O teste da PB – Leitura, apesar de possuir elementos de ordem técnica e conceitual para a constituição de uma avaliação educacional e de seu amplo uso por professores dos anos iniciais do ensino fundamental, apresentava limitações que poderiam ser superadas por testes adaptados aos perfis de aprendizagem dos estudantes e com resultados mais fidedignos para apoiar as decisões pedagógicas de professores(as) e gestores(as) escolares. Assim, buscou-se responder à questão: "É possível construir um TAI para a versão impressa da PB – Leitura que seja ponto de apoio para professores(a) na avaliação de alunos(as) dos anos iniciais do ensino fundamental?". Para a construção dessa ferramenta TAI da PB – Leitura foi necessário articular engenheiros de *softwares*, elaboradores de testes, pesquisadores e profissionais da educação de 15 escolas públicas do município de São Paulo. Para que pudessem participar da construção da ferramenta e da validação dos resultados obtidos, foi realizada formação de professores(as) e gestores(as) educacionais sobre medida educacional, leitura e avaliação. Após a verificação de que os aspectos psicométricos dos itens da versão impressa poderiam ser mantidos para a versão informatizada, o TAI da PB – Leitura foi aplicado e os resultados indicaram que ele permitiu testes personalizados aos domínios dos(as) alunos(as), mais rápidos e de menor comprimento, sem prejuízo da precisão. Por apresentar resultados embasados em uma escala com importante interpretação pedagógica, o TAI da PB – Leitura se revelou capaz de apoiar a prática avaliativa de professores(as) e gestores(as) e o trabalho pedagógico na alfabetização e no letramento inicial. Esse apoio foi potencializado com o acréscimo de uma regra ao critério de parada do TAI, utilizada em testes que visam a classificação do respondente em níveis de resultado. Verificou-se também a necessidade de aprofundar as investigações sobre: a formação de

professores(as) na temática da medida e avaliação; a ampliação do banco de itens, com a finalidade de controle de taxas de exposição e balanceamento de conteúdo, e a produção de relatórios pedagógicos.

Palavras-chave: Provinha Brasil. Avaliação educacional. Teste Adaptativo Informatizado (TAI). Proficiência em leitura. Alfabetização.

ABSTRACT

CATALANI, Érica Maria Toledo. **Computerized Adaptive Test of Provinha Brasil**: the construction of a supportive instrument for teachers and school administrators. 2019. 282 p. Thesis (Doctorate in Education) – Faculdade de Educação, Universidade de São Paulo, São Paulo, 2019.

This thesis results from a project of construction of a Computerized Adaptive Test (CAT) for the paper and pencil version of Provinha Brasil (PB), focused on the assessment of proficiency in reading. The PB – Reading test, despite having technical and conceptual elements for the constitution of an educational assessment and its wide use by teachers of the initial years of elementary school, presented limitations that could be overcome by tests adapted to the learning styles of students and with much more reliable outcomes to support the pedagogical decisions of teachers and school administrators. Thus, it was sought to answer the question: "Is it possible to create a CAT for the printed version of PB – Reading test which would be a base of assistance for teachers in the assessment of students in the initial years of elementary education?" For the creation of this CAT tool from PB – Reading test it was necessary to articulate software engineers, test designers, researchers and education professionals from 15 public schools from São Paulo city. In order to take part in the creation of the tool and the validation of the achieved results, it was made teachers and educational managers training on educational measures, reading and assessment. After verifying that the psychometric aspects of the printed version items could be kept for the computerized version, the PB – Reading CAT was applied and the results indicated that it allowed customized testing to the student's domains, faster and of smaller length, without prejudice of the precision. Based on a scale with an important pedagogical interpretation, the PB – Reading CAT was able to support the assessment practice of teachers and managers and the pedagogical work in literacy and initial literacy. This support was strengthened by adding a rule to the CAT stopping criterion, used in tests that aim to classify the respondent into outcome levels. There was also a need to deepen the research on: teacher training in the subject of measurement and assessment; the expansion of the item base, for the purpose of controlling exposure rates and content balancing, and the production of pedagogical reports.

Keywords: Provinha Brasil. Educational assessment. Computerized Adaptive Testing (CAT). Reading Diagnostics. Literacy.

LISTA DE FIGURAS

LISTA DE EQUAÇÕES

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
ACA	Ambiente Computacional de Aprendizagem
ANA	Avaliação Nacional da Alfabetização
AVASB	*Armed Services Vocational Aptitude Battery*
BI	Banco de Itens
Capes	Comissão de Aperfeiçoamento de Pessoal de Nível Superior
CAT	*Computerized Adaptive Testing*
CCI	Curva Característica do Item
CBT	*Computer-Based Test*
CEU Emef	Escola Municipal de Ensino Fundamental do CEU
CEU	Centro Educacional Unificado
COD.INEP	Código atribuído para o item pelo Inep
Daeb	Diretoria de Avaliação da Educação Básica
DRE	Diretoria Regional de Educação
DU	*Design* Universal
EAP	Estimador bayesiano da média *a posteriori*
Encceja	Exame Nacional para Certificação de Competências de Jovens e Adultos
Enem	Exame Nacional do Ensino Médio
Emef	Escola Municipal de Ensino Fundamental
EMEFM	Escola Municipal de Ensino Fundamental e Médio
Feusp	Faculdade de Educação da Universidade de São Paulo
ICMC	Instituto de Ciências Matemáticas e de Computação
Gepave	Grupo de Estudos e Pesquisas em Avaliação Educacional
GB	*Gigabyte*
GBA	*Game-Based Assessment*
GMAT	*Graduate Management Admission Tests*
GRE	*Graduate Record Exam*
Inep	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
kB	Kilobyte

MEC	Ministério da Educação
MB	*Megabytes*
MFI	Máxima Informação de Fisher
ML1P	Modelo logístico de um parâmetro
MST	Testes de Múltiplos Estágios
NCLEX	National Council of State Boards of Nursing
MH	Método de Mantel-Haenszel
NTA	Núcleo Técnico de Avaliação
PB	Provinha Brasil
PB – Leitura	Teste para diagnóstico da Leitura da Provinha Brasil
Pnaic	Pacto Nacional pela Alfabetização na Idade Certa
PNE	Plano Nacional da Educação
PPT	*Paper-and-Pencil Test* [Teste em papel e lápis]
PRODOC	Programa de Apoio a Projetos Institucionais com a Participação de Recém-Doutores
PROMIS	Sistema de Informação de Medição de Resultados do Instituto Nacional de Saúde dos EUA
RAM	*Random Access Memory*
RME/SP	Rede Municipal de Ensino de São Paulo
Saeb	Sistema de Avaliação da Educação Básica
SME/SP	Secretaria Municipal de Educação de São Paulo
TAC	Teste(s) Adaptativo(s) Computadorizado(s)
TAI	Teste(s) Adaptativo(s) Informatizado(s)
TAEPI	Testes Adaptativos para o Exame de Proficiência em Inglês
TBC	Teste(s) Baseado(s) em Computador(es)
TCM	Teoria Clássica das Medidas
TCT	Teoria Clássica dos Testes
TIC	Tecnologia da Informação e Comunicação
TOEFL	*Test of English as a Foreign Language*
TRI	Teoria da Resposta ao Item
TTT	Teoria Tradicional dos Testes
UE	Unidade(s) Educacional(is)

Unesco	Organização das Nações Unidas para a Educação, a Ciência e a Cultura
USP	Universidade de São Paulo

LISTA DE SÍMBOLOS

θ	theta verdadeiro, traço latente ou proficiência
$\hat{\theta}$	theta estimado
e	número de Euler (e = 2,718...)
L	função de verossimilhança
ln L	função log-verossimilhança (*log likelihood function*)
T	pontuação verdadeira (*true score*)
$I(\theta)$	função de informação total
$I_j(\theta)$	função de informação do item j
$\prod$	produtória
S_e	erro padrão
D	constante da função logística da TRI
x_i	resposta dada ao item no momento i
$\propto$	proporcional

SUMÁRIO

1 INTRODUÇÃO

Meu percurso profissional como professora de Matemática, coordenadora pedagógica, supervisora escolar e especialista em avaliação na Secretaria Municipal de Educação de São Paulo contribuiu para que eu passasse a integrar, em 2013, o Grupo de Estudos e Pesquisas em Avaliação Educacional (Gepave)[1], vinculado à Faculdade de Educação da Universidade de São Paulo (Feusp).

Os estudos realizados no Gepave foram cruciais para a percepção de que as questões sobre avaliação educacional ganharam impulso com a implantação de avaliações externas no Brasil, na década de 1990, especialmente com a criação do Sistema de Avaliação da Educação Básica (Saeb)[2] pelo Ministério da Educação (MEC), ampliando o debate de ordem conceitual e prático já travado no âmbito das avaliações internas.

De um lado, a avaliação externa recebia críticas por se basear em processos de medida educacional, cujos resultados eram colocados em identidade com a qualidade da educação, por desconsiderar a presença de gestores(as) e professores(as) na sua consecução e por ser sobreposta à avaliação realizada por professores(as) e escolas.

De outro, as taxas de reprovação, abandono e evasão, conservadas em patamares pouco satisfatórios, bem como estudos sobre as práticas avaliativas desencadeadas no interior das escolas revelavam que a avaliação de caráter formativo e de rompimento com a exclusão, expressos em determinadas práticas classificatórias e de seletividade dos(as) alunos(as) – aspectos amplamente tratados na literatura (Cf. HOFFMANN, 2003; LUCKESI, 2008) –, estava longe de ser alcançada.

Costumeiramente, o debate ainda colocava em polos antagônicos avaliação interna e externa, somativa e formativa, objetividade e subjetividade, além de tratar os processos de medição e de avaliação como sinônimos.

---

1 Grupo de pesquisa vinculado à Faculdade de Educação da Universidade de São Paulo (Feusp), coordenado pelo Professor Dr. Ocimar Munhoz Alavarse.

2 O Saeb é um conjunto de instrumentos que permite a produção e a disseminação de evidências, estatísticas, avaliações e estudos a respeito da qualidade das etapas que compõem a educação básica no Brasil (BRASIL, 2018).

A avaliação somativa, por ser realizada ao final de um período de escolarização, era frequentemente relacionada à avaliação externa; já a avaliação formativa, por fornecer informação para a atuação do(a) professor(a), à avaliação interna – embora essas denominações na prática extrapolassem contornos tão definitivos, pois nem toda avaliação interna necessariamente apresenta o caráter formativo e nem toda avaliação externa proporciona informações unicamente referentes ao final do processo.

Nesse contexto, a Provinha Brasil (PB), um instrumento padronizado e disponibilizado para todo o território brasileiro na versão impressa pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep)[3], constitui um exemplo típico dessa extrapolação. Embora a PB mantenha fortes características de uma avaliação externa – por oferecer um teste padronizado, construído com o suporte de avaliadores externos e agregado a uma escala de proficiência, elaborada sob os preceitos da medida educacional, procurando conferir maior validade e fidedignidade aos resultados –, apresenta também características que satisfazem os aspectos práticos das avaliações internas – por submeter a aplicação do teste, a apuração dos acertos e a análise dos resultados aos(às) próprios(as) educadores(as) das escolas; por possibilitar a utilização desses resultados para intervenções pedagógicas no processo de alfabetização e letramento inicial em Língua Portuguesa, no aspecto da leitura, e em Matemática, no aspecto da resolução de problemas; e por permitir diagnóstico rápido para acompanhamento dos(as) gestores(as) das escolas. O teste impresso, construído para diagnóstico da leitura da PB, doravante tratado por PB – Leitura, foi proposto pelo Inep de 2008 a 2016[4] para ser aplicado em todos os municípios do Brasil. A utilização desse teste, no contexto das escolas, intensificou-se ainda mais quando passou a fazer parte do monitoramento das ações do Pacto Nacional pela Alfabetização na Idade Certa (Pnaic), consagrado em legislação no ano de 2012[5], conferindo-lhe relevância política além da pedagógica.

Nos estudos realizados, sobretudo aqueles vinculados à pesquisa da qual deriva este trabalho, com vistas às tensões entre a avaliação interna e a externa, distinguiram-se

---

[3] Autarquia vinculada ao Ministério da Educação.
[4] Em 8 de agosto de 2016, foi noticiada a suspensão da Provinha Brasil, que passaria a ser disponibilizada apenas em arquivo digital (TOKARNIA, 2018).
[5] Pacto assumido nacionalmente para garantia dos direitos à alfabetização e ao letramento inicial.

três núcleos, considerados primordiais para configurar o problema da pesquisa. O primeiro esteve focado na medida educacional, visto que as principais avaliações externas brasileiras adotavam a versão impressa de testes, também denominada papel e lápis, e apresentavam o problema do erro de medida que, embora em patamares controlados, poderiam ser melhorados. Uma alternativa para o aperfeiçoamento dos processos de avaliação, particularmente em relação aos seus instrumentos e procedimentos, sobre os quais nos debruçamos e que se constituem no objeto desta tese são os Testes Adaptativos Informatizados (TAI)[6] (Cf. ALAVARSE; MELO, 2013a, 2013b; ALAVARSE; CATALANI, 2016a, 2016b; KLEIN, 2013). A aferição realizada nas avaliações externas para fins de dimensionar o conhecimento ou a competência que se consistiria no objeto dessas avaliações, embora pautadas na utilização de sofisticadas análises por meio de modelos da Teoria da Resposta ao Item (TRI) para conferir maior grau de validade e fidedignidade aos testes, ainda apresentam os limites típicos dos instrumentos de avaliação administrados com o uso de papel e lápis, nos quais a estimação da proficiência ocorre por meio de um conjunto fixo de itens submetido a todos os participantes independentemente da variação da proficiência de seus respondentes. Essa estimação fica principalmente prejudicada para os participantes que se encontram com proficiências muito baixas ou muito altas no espectro da escala de medida (MELO, 2017). Em face dessa situação, passamos a estudar os TAI, vislumbrando com seu emprego uma melhor aferição da proficiência para esses(as) respondentes, pois, de modo geral, esse tipo de teste, incluindo seus dispositivos eletrônicos de administração, identifica o domínio do(a) respondente ao verificar como reage aos primeiros itens, adequando o nível de dificuldade dos próximos itens que serão administrados. Esse procedimento permite que o(a) respondente receba um teste com itens ajustados aos seus domínios, elemento que melhora a precisão da estimativa. A literatura ainda indicava que o TAI poderia conferir maior rapidez na obtenção dos resultados, ao substituir pela automação o processo manual de contagem dos acertos e alocação dos(as) respondentes nos níveis da escala.

O segundo núcleo da tensão entre avaliação externa e interna mostrava a necessidade de trilhar um caminho coincidente com a percepção de Nevo (1997) de

---

[6] Em inglês, o termo é *Computerized Adaptive Testing* ou *Computering Adaptive Testing* (CAT).

que a avaliação externa e a interna apresentavam características diferentes e importantes para o processo de avaliação e que, ao contrário de se estabelecerem como antagônicas, poderiam ser colocadas em *diálogo*, campo fértil para desenvolver a reflexão sobre a prática de avaliadores das escolas, fundamentalmente quanto aos processos que implementam.

O terceiro núcleo foi demarcado pela importância da avaliação da proficiência em leitura dos(as) alunos(as), sobretudo nos anos iniciais do ensino fundamental, notadamente com a preocupação de que a autonomia leitora das crianças constitui um dos objetivos essenciais do processo de alfabetização inicial e os(as) educadores(as) destinam grandes esforços ao processo de alfabetização e letramento inicial nos anos iniciais do ensino fundamental, pois, entre outros elementos, nesse período ocorre o estudo formal sobre a natureza e as características do sistema de escrita e o desenvolvimento de estratégias de leitura. Dado o lugar de relevo da autonomia leitora, além da preocupação com as concepções de alfabetização dos(as) educadores(as), o modo como avaliam a proficiência leitora nessa etapa escolar pode representar um entrave para essa apropriação, considerando que essa avaliação proporciona o apoio para a intervenção pedagógica.

Assim, os cuidados com procedimentos avaliativos, na perspectiva formativa e da construção de um projeto pedagógico inclusivo, aliados às características do teste da PB – Leitura, sua larga utilização no processo inicial de alfabetização e letramento inicial, ofereciam condições favoráveis para uma investigação, a despeito de considerar as polêmicas sobre suas abordagens pedagógicas (Cf. MICOTTI, 2013) e sua avaliação (Cf. ESTEBAN, 2009; MORAIS, 2012). Quanto às condições favoráveis, além da importância político-pedagógica da PB, sublinhavam-se os elementos característicos de uma avaliação educacional, presentes na sua concepção, quais sejam: a) uma matriz de avaliação ou de referência, na qual se estabelece o objeto de avaliação; b) banco de questões/itens[7] que são elaborados por representantes de educadores(as) de diferentes regiões do país, revisados, previamente testados e com tratamento psicométrico; c) itens que apresentam boa cobertura da matriz e parametrizados em uma escala de proficiência; d) escala de proficiência, subdividida

[7] Item consiste na unidade básica de um instrumento de coleta de dados, que pode ser uma prova, um questionário etc. Nos testes educacionais, item pode ser considerado sinônimo de questão, termo mais popular e utilizado com frequência nas escolas.

em cinco níveis, que adota procedimentos de medida para dimensionar diferenças de desempenho e que possibilita critérios comuns para o julgamento dos resultados; e) interpretação pedagógica para os cinco níveis, facultando ações pedagógicas em decorrências desses níveis; e f) uma fundamentação teórica documentada de todos esses elementos.

Como um elemento extremamente importante depreendido do tratamento da literatura, para a elaboração de um TAI experimental, haveria a necessidade de uma equipe multidisciplinar de especialistas em avaliação, desenvolvimento de *softwares* e plataformas, programadores(as) e educadores(as); recursos financeiros de grande vulto para construção de banco de itens e para os dispositivos eletrônicos usados na administração do teste; e o envolvimento de muitos(as) respondentes, o que nos levou à pesquisa do desenvolvimento do projeto intitulado TAI da PB – Leitura, sigla usada para se referir ao projeto que visou construir uma versão adaptativa e informatizada do teste diagnóstico da leitura da PB e que será foco do presente trabalho.

Com a identificação desses três núcleos e a definição do acompanhamento desse projeto delineou-se a questão de pesquisa: “É possível construir um TAI para a versão impressa da PB – Leitura que seja ponto de apoio para avaliação de alunos(as) dos anos iniciais do ensino fundamental?”.

Mesmo considerando toda a sorte de obstáculos que o desenvolvimento da pesquisa teria e que seus resultados se constituiriam em um ensaio inicial, decidiu-se por sua realização, contando com o apoio de colegas do Gepave, pela relevância política e pedagógica de um TAI voltado para a leitura, ainda mais pela ausência de algo similar nos levantamentos de referências conduzidos. Salienta-se que sem os aportes recebidos do Gepave, com a disponibilização de pesquisadores especialistas em alfabetização e avaliação; da Secretaria Municipal de Educação de São Paulo (SME/SP), com a mobilização de servidores, estagiários(as) e consultores via Prodoc/Unesco[8], para o desenvolvimento da plataforma informatizada e do algoritmo para o projeto TAI da PB – Leitura; e do Inep, que disponibilizou relatórios técnicos sobre a Provinha Brasil, esse trabalho não poderia ser desenvolvido. Cabe ainda destacar a contribuição do pesquisador Ailton Carlos Santos na coordenação da

[8] Projeto de cooperação técnica articulado pela Organização das Nações Unidas para a Educação, a Ciência e a Cultura (Unesco) do Brasil.

formação em alfabetização e letramento inicial, realizada durante o projeto. Essa formação envolveu gestores(as) e professores(as) de quinze unidades educacionais da Rede Municipal de Ensino de São Paulo (RME/SP) e tinha como princípio a participação dos(as) educadores(as) no processo de construção do instrumento, vislumbrando que ele seja um apoio fidedigno e rápido para a avaliação da leitura na faixa etária em questão.

O trabalho foi estruturado em quatro capítulos, além desta Introdução. O Capítulo 2 aborda os principais conflitos no campo educacional entre avaliação e medição, enfatizando suas implicações para os processos de ensino e de aprendizagem. Argumenta-se de um lado que, embora medir e avaliar não sejam sinônimos, o julgamento, aspecto central do processo de avaliação, pode ser assentado em uma medida; de outro, que o teste da PB – Leitura, a despeito de ter características de um processo de avaliação, apresenta limitações que podem ser superadas por um TAI. No Capítulo 3, são proporcionadas as características e a revisão da literatura sobre os TAI. A descrição das etapas de aplicação do TAI da PB – Leitura bem como as reflexões sobre os resultados encontram-se no Capítulo 4. O Capítulo 5 traz as conclusões, sendo que os resultados indicaram que foi possível construir uma versão informatizada para a PB – Leitura que proporcionou: personalização dos testes aos conhecimentos dos(as) alunos(as), maior economia de tempo na aplicação e menor comprimento do teste, sem prejudicar a fidedignidade. O TAI construído possibilitou a diferenciação dos níveis de proficiência dos(as) alunos(as), característica essencial para apoiar o processo de avaliação e as intervenções pedagógicas necessárias ao processo de alfabetização e letramento inicial, contemplando a perspectiva formativa da avaliação. Nesse capítulo, também são identificadas as limitações do trabalho com a apresentação de perspectivas de estudos futuros em relação à necessidade de ampliação do banco de itens, do controle da taxa de exposição e do balanceamento do conteúdo, de ampliação do estudo sobre a formação de educadores em torno da temática da medida educacional e sobre a apresentação de relatórios de resultados.

2 AVALIAÇÃO, MEDIÇÃO, ENSINO E APRENDIZAGEM

O TAI é uma ferramenta concebida para oferecer informações sobre um domínio cognitivo, integrando uma etapa do processo de avaliação educacional. Contudo, essa ferramenta pode servir de apoio para a avaliação educacional estando baseada em um processo de medida?

Essa pergunta tem sentido porque, além de estreita, a relação entre avaliação e medição é vista de forma distorcida e ponderamos que se trata de um aspecto pouco explorado na literatura do campo da avaliação educacional, não bastando a sinonímia que amiúde se estabelece entre essas operações, via de regra considerando que basta medir – estimar – uma proficiência para que a avaliação tenha se completado. Neste capítulo, pretende-se abordar as principais tensões entre avaliação e medição no campo educacional, procurando, ainda, demarcar suas implicações para os processos de ensino e de aprendizagem.

2.1 As tensões no campo da avaliação educacional

Sem dúvida, a avaliação educacional é um dos conceitos considerados mais complexos, tendo em vista que se relaciona com todos os outros elementos intervenientes do processo educativo, como currículo, processos de ensino e de aprendizagem e formação de professores(as) (ÁLVAREZ MENDEZ; 2002).

Essa complexidade e a constatação de que esse tema tem sido relativamente negligenciado na formação inicial dos(as) professores(as) contribuíram para fomentar diversas polêmicas no campo da avaliação educacional no Brasil (Cf. ALAVARSE, 2015; GATTI et al., 2010).

As polêmicas se intensificaram principalmente após a instituição, pelo Ministério da Educação (MEC), de avalições externas que utilizam testes de aferição da proficiência em leitura e escrita e resolução de problemas. A realização periódica de uma

avaliação educacional censitária, com frequência bianual[9], de um lado possibilitou a construção de indicador importante[10] para acompanhar o desenvolvimento da educação básica, a despeito da definição de uma proficiência para a escola com base no cálculo da média das proficiências dos(as) alunos(as) (MELO, 2017); de outro, trouxe a preocupação com a questão da redução da qualidade ao indicador que, entre outros aspectos, limita-se ao concentrar-se na aferição de uma parcela dos conhecimentos e das competências desenvolvidos na escola e ao eleger a avaliação como núcleo da política educacional, deixando de lado o currículo e a formação de professores(as). Os aspectos negativos atribuídos às avaliações externas acentuaram-se com a sobreposição dessas avaliações, visto que os governos estaduais e municipais também passaram a realizar as suas, muitas vezes vinculando os resultados aferidos às bonificações salariais (Cf. ALAVARSE, 2015; GATTI, 2014).

Embora as tensões tenham sido acirradas com a chegada da avaliação externa, não significa que antes havia tranquilidade entre os pesquisadores do campo. A relação da avaliação com o ensino e a aprendizagem já era conflituosa, especialmente, por não estar alinhada ao propósito de garantir o sucesso da aprendizagem, sua função precípua (LUCKESI, 2018). Em relatos de experiências sobre a avaliação educacional, não era difícil encontrar situações em que a avaliação era usada para garantir o controle do comportamento na aula, a motivação para o estudo de uma dada temática ou a frequência dele(a) na aula. O medo e a arbitrariedade foram aspectos constantemente atribuídos aos momentos de prova (HOFFMANN, 2003), que, por usa vez, em muitas situações, continua sendo identificada como sinônimo de avaliação, embora seja de fato um dos instrumentos de coleta de informações que integra o processo de avaliação.

Como outra face dessas considerações no terreno da avaliação educacional, temos que o processo de medição, no âmbito educacional, que fundamenta a construção dos testes, tem sido um dilema entre pesquisadores envolvidos nesse debate,

---

[9] No início dos anos 1990, foi criado o Sistema de Avaliação da Educação Básica (Saeb), desdobrado em 2005 na Avaliação Nacional do Rendimento Escolar (Anresc), denominada Prova Brasil, e na Avaliação Nacional da Educação Básica (Aneb). Em 2013, o Saeb desdobrou-se também na Avaliação Nacional da Alfabetização (ANA), cuja periodicidade anual, definida inicialmente, não se concretizou.

[10] Índice de Desenvolvimento da Educação Básica (Ideb), instituído em 2007 e cujo cálculo agrega a média das proficiências em Língua Portuguesa e Matemática e as taxas de aprovação para os resultados da Prova Brasil desde 2005.

constituindo um espectro de posições com extremos nos quais encontramos desde os que negam peremptoriamente a possibilidade, e a própria necessidade, de medição de atributos cognitivos individuais, como seria a proficiência em leitura, até aqueles que parecem se limitar a medições desses traços para que a avaliação esteja concluída. A isso se agrega a ausência, quase absoluta, de formação dos profissionais da educação para lidar com teorias e modelos para a medição e o dimensionamento dos domínios dos(as) estudantes, como se consolidou no campo da psicometria em seus desdobramentos para a educação, donde se poderia considerar a existência de uma edumetria.

Quanto ao caráter somativo e formativo das avaliações, é preciso esclarecer que a avaliação somativa corresponde à síntese avaliativa realizada ao final de processos educativos e, no âmbito da instituição educacional, tem sido relacionada à atribuição de notas, conceitos ou descrições sobre a situação final de desempenho de alunos(as); já a avaliação formativa, a partir de seus resultados, deve orientar a ação do(a) professor(a), acarretando a investigação das causas para os resultados não satisfatórios encontrados e, eventualmente, o replanejamento das práticas pedagógicas necessárias ao processo escolar voltado para o sucesso da aprendizagem de todos. (Cf. ÁLVAREZ MENDEZ, 2002; FERNANDES, 2009; FERNANDES; FREITAS, 2007; RUSSELL, 2010).

É comum relacionar a avaliação somativa, mais voltada para observação de aspectos abrangentes do conhecimento e relativa ao conjunto de conhecimentos consolidados em período mais alongado de tempo, à avaliação externa; por seu turno, a avaliação formativa, mais conectada aos conhecimentos da sala de aula, é considerada mais relacionada à avaliação interna. No entanto, esses contornos não são tão definitivos na prática, pois nem toda avaliação interna é realizada para orientar a ação do(a) professor(a), deixando de ter o caráter formativo. Além disso, é possível que os resultados de uma avaliação externa proporcionem informação útil para replanejamento do professor(a).

Cabe também apresentar elucidações sobre as dimensões interna e externa da avaliação educacional (Cf. ALAVARSE, 2013; BLOOM; HASTINGS; MADAUS, 1983; FERNANDES, 2009; FREITAS, 2007). A interna corresponde à avaliação feita pelo(a) professor(a) como parte do seu fazer pedagógico e a externa é a avaliação feita por agentes externos à escola. Na avaliação interna, a coleta de informações ocorre com

maior frequência/periodicidade, dada a proximidade do(a) avaliador(a), e existe menor abrangência quanto aos atributos considerados para a descrição e o julgamento.

Os instrumentos de coleta na avaliação interna, quando não planejados, podem priorizar aspectos que são alheios ao conhecimento que se quer descrever como, por exemplo, a coleta de informações sobre um conhecimento matemático ser influenciada/distorcida pelo conhecimento de língua portuguesa (interpretação de texto), comprometendo a validade[11] das informações coletadas. A ausência de padronização de instrumentos e critérios de análise, nas avaliações internas, são aspectos que podem gerar divergências nefastas no processo de avaliação de estudantes que estão vinculados a uma mesma instituição. Outro aspecto sobre a avaliação interna, já apontado, refere-se à ausência de transparência quanto aos critérios de análise e atributos envolvidos na avaliação, aspectos que corroboram a sensação de arbitrariedade do processo. O exposto revela que essa dimensão apresenta vantagens, mas também limitações.

A avaliação externa, por sua vez, estrutura-se para evidenciar maior transparência na sua implementação, oferecendo: as especificações do atributo avaliado, com antecedência; critérios comuns de análise dos resultados, que são equivalentes para todos(as); verificação da validade das informações coletadas, ao analisar a quantidade de fatores que incidem nas respostas observadas; a padronização das condições de aplicação; e a precisão na aferição das proficiências. Contudo, em uma avaliação externa faz-se uma única coleta por ano ou por períodos maiores, o que provoca a desconfiança de educadores(as) e diretores(as) sobre as inferências fornecidas. Por serem geralmente realizadas em larga escala, fundamentadas em testes de papel e lápis, com ênfase em alguns conhecimentos e em itens de múltipla escolha, essas avaliações são consideradas restritas na abrangência das informações que conseguem coletar para análise, de modo que essa dimensão também apresenta vantagens e limitações.

Tendo por princípio colocar a escola como centro das ações da política educacional, Nevo (1997) aponta a necessidade do diálogo entre as duas dimensões – a externa e interna –, que ele identifica como tipos, mostrando que ambas apresentam

---

[11] Validade é a correspondência entre o objeto (um construto ou atributo específico) que se quer medir e sua/seu manifestação/comportamento (Cf. PASQUALI, 2013).

características diferentes e complementares, importantes para o processo de avaliação.

Sem focar nos antagonismos, mas na criação de um discurso mais construtivo para a avaliação educacional, Nevo propõe combinar esses dois tipos de avaliação, considerando que a avaliação interna tende a ser mais subjetiva e enviesada que a avaliação externa e esta, por sua vez, promove disputas e condutas defensivas por parte de educadores(as) e gestores(as) e se baseia em coletas de informação consideradas restritas, em termos da abrangência do processo de ensino e aprendizagem.

Além disso, o autor propõe que as escolas devam desenvolver suas capacidades de avaliação com o diálogo entre avaliadores internos e externos. Aponta ainda que esse diálogo é essencial para a compreensão das necessidades e percepções de professores(as) e diretores(as) sobre o modo como interpretam e utilizam os resultados de um processo avaliativo. Na experiência de trabalho com as escolas, o autor defende que o diálogo e a formação são primordiais para que professores(as) e diretores(as) possam compreender e colocar em prática etapas envolvidas na medição educacional, implementando-as em processos internos de avaliação. Em contrapartida, os avaliadores externos, além de contribuir nessa formação, podem combinar os resultados da avaliação externa aos elementos do contexto educacional.

Em um caminho paralelo, Russell (2010) mostrou que, no geral, as pesquisas sobre o uso de tecnologias baseadas em computador na avaliação educacional concentraram-se em fornecer testes educacionais para avaliações somativas e que só recentemente um pequeno, mas crescente, corpo de pesquisa começou a investigar o uso de computadores para avaliação formativa. Para o autor, a avaliação formativa é o processo de coleta que permite obter informações sobre o conhecimento do(a) aluno(a) antes ou durante uma intervenção pedagógica. Os propósitos desse tipo de avaliação se dividem: de um lado visam informar o gestor da ação pedagógica para que possa definir o apoio pedagógico que seja mais proveitoso; de outro, informar os(as) alunos(as) sobre aos avanços e as dificuldades, fornecendo *feedback*s adequados (Cf. CARDINET, 1993). Ademais, para Russell, a diferença nas duas dimensões é observada tanto para a formulação do teste quanto na lacuna de tempo entre a coleta e o retorno dos resultados. Na avaliação formativa, considerada uma avaliação mediadora, as informações coletadas devem estar mais próximas e

alinhadas ao trabalho da aula e o retorno deve ser imediato, a fim de informar professores(as) e educandos(as). Em vez disso, na avaliação somativa serão identificados os tópicos e as habilidades mais abrangentes dominados pelos(as) alunos(as), com o intuito de identificar se precisam de mais desenvolvimento, ao considerar que existe uma lacuna entre o recebimento das informações somativas e a próxima oportunidade de desenvolver mais conhecimentos e habilidades específicas, impedindo o uso de testes somativos como forma de avaliação formativa.

A opção pelo teste da PB – Leitura ocorreu pela percepção de um certo alinhamento entre o teste na versão papel e lápis e a prática de sala de aula, consubstanciando o diálogo entre a avaliação externa e a avaliação interna. Sem isso, o desenvolvimento da ferramenta TAI da PB – Leitura poderia fracassar no quesito apoio ao processo de alfabetização e letramento iniciais. Esse alinhamento ganhou contornos ainda maiores no grupo de professores(as) e gestores(as) envolvidos(as) no projeto, em razão da formação, que proporcionou o diálogo mediado por avaliadores(as), revelando tanto os limites como as potencialidades de itens e de testes. Entre os limites fortemente observados pelos(as) educadores(as), estava a ausência de diagnóstico da escrita, altamente valorizado por eles(as). Itens que avaliam a hipótese de escrita, conforme Brasil (2012b), foram eliminados dos instrumentos da PB – Leitura após a primeira edição (2008), por problemas metodológicos, ou seja, era difícil garantir a padronização da análise de um item de resposta construída, já que eram os(as) próprios(as) professores(as)/aplicadores(as) que realizavam essa análise. Esse problema, embora não contornado no TAI da PB – Leitura desenvolvido, estaria superado ao se utilizar a pontuação informatizada, possível em ferramentas TAI. (Cf. BEJAR, 2011; WILLIAMSON; MISLEVY; BEJAR, 2006).

2.2 Medição e avaliação

A medida foi e ainda é muito criticada – questionada – no âmbito da avaliação educacional. Diferentes autores se dedicaram em dizer que avaliar não é medir, mas poucos conseguiram diferenciar medida de avaliação ou caracterizar o papel da medida no processo de avaliação.

Se medir não é avaliar, como articular medida e avaliação?

Para responder a essa pergunta, um fator essencial do projeto estava em detalhar as diferenças entre o significado da medida e da avaliação com um diferencial: apontar os limites e as possibilidades de uma medida educacional e como ela pode embasar a avaliação. Desse modo, reafirmou-se que “medir não é avaliar”, mas elucidou-se que a avaliação pode estar alicerçada em uma medida (CARDINET, 1993; LUCKESI, 2018; LUKAS MUJIKA; SANTIAGO ETXEBARRÍA, 2009).

A avaliação educacional é tarefa complexa, pois, entre outras características, inter-relaciona-se, entrelaça e, às vezes, funde-se com duas outras temáticas de relevo no âmbito educacional: o currículo e a formação docente. Por esse motivo, para aclarar o significado da avaliação educacional e sua diferença com o aspecto da medida, partimos de uma definição que agrega a maioria dos elementos presentes em definições disponíveis na literatura e é dada por Lukas Mujika e Santiago Etxebarría, (2009, p. 91-92):

> A avaliação é o processo de identificação, levantamento e análise de informação relevante de um objeto educacional – que poderá ser quantitativa ou qualitativa –, de forma sistemática, rigorosa, planificada, dirigida, objetiva, crível, fidedigna e válida para emitir juízo de valor baseado em critérios e referências preestabelecidos para determinar o valor e o mérito desse objeto a fim de tomar decisões que ajudem a otimizá-lo. (tradução nossa)

Essa definição não pode ser isolada da perspectiva histórica no campo da avaliação educacional, pois esse campo passou por modificações ao longo do tempo[12].

De todo modo, possibilita visualizar as diferentes etapas do processo de avaliação que, se não estão distintas na ação cotidiana, podem ocorrer sem a intencionalidade necessária, podendo comprometer a transparência do processo. Vale destacar que a definição não coloca os processos de medida como desfavoráveis à avaliação, mas podem incorporá-los como possíveis suportes para o julgamento – elemento fulcral da definição de avaliação.

---

[12] Como esse não é o foco principal desta investigação, para obter mais detalhes sobre os diferentes modelos e abordagens, estabelecidos a partir da observação das vantagens e limitações, conferir Guba e Lincoln (1989) e Fernandes (2009).

As recentes considerações epistemológicas de Luckesi (2018) sobre o ato de avaliar, definido como o ato de investigar a qualidade de uma realidade, também estabelecem passos metodológicos da investigação avaliativa que se relacionam à definição apresentada.

Um componente fundamental da avaliação, expresso nessa definição, é a identificação da informação relevante sobre um "objeto educacional". O objeto que será investigado no processo de avaliação deve ser delimitado, com a identificação dos conhecimentos e das habilidades e das variáveis que conjuntamente configuram os dados essenciais a serem coletados. Nesse aspecto, Luckesi (2018, p. 47) ressalta a ausência dessa prática nas escolas.

É habitual, nos processos de avaliação nas escolas, eleger os conhecimentos e as competências de alunos(as) como objetos únicos. No entanto, isso é considerado uma limitação, visto que é necessário combinar essas informações com as de outros objetos educacionais, como o currículo, os planos de ensino, os projetos e outros componentes que perpassam o trabalho escolar na avaliação (FERNANDES, 2009; LUCKESI, 2018; NEVO, 1997).

No processo de medição que fundamenta os testes, a delimitação do objeto é representada por meio de uma matriz de especificações ou, no caso da PB, por meio da matriz de avaliação, na qual se estabelece a taxonomia dos conhecimentos e habilidades que serão objeto da avaliação, visando à abrangência ou à cobertura dos conhecimentos e das habilidades no instrumento de coleta.

A depender da própria natureza do objeto a ser observado, pode-se adotar a observação direta da performance do educando ou o recurso a tarefas em que as performances – respostas às tarefas propostas – podem ser interpretadas posteriormente, como nos testes escritos ou nas provas com itens de múltipla escolha. Como exemplo, podemos tomar a habilidade de cálculo mental em Matemática, aspecto que só pode ser inferido a partir de uma performance do(a) aluno(a), ao contrário, a habilidade do cálculo de operações, que pode ser inferida por meio de um teste, cujas respostas serão processadas posteriormente.

Embora os contornos do objeto avaliado no TAI da PB – Leitura já sejam definidos para o teste em papel e lápis, a reflexão sobre esse objeto, referente à competência leitora a ser diagnosticada com um TAI para estudantes dos anos iniciais, constituiu

um grande desafio no debate com professores(as) e gestores(as), pois a proficiência aferida no processo de alfabetização e letramento compreende diferentes aspectos do conhecimento estudados na Língua Portuguesa.

Nos documentos que acompanham os testes da PB – Leitura em papel e lápis, esses aspectos são chamados de eixos e estão definidos na Matriz de Referência para Avaliação da Alfabetização e do Letramento Inicial. O documento Brasil (2016a) estabelece como habilidades imprescindíveis para o desenvolvimento da alfabetização e do letramento inicial as que podem ser agrupadas em torno de cinco eixos: 1) apropriação do sistema de escrita, 2) leitura, 3) escrita, 4) compreensão e valorização da cultura escrita e 5) desenvolvimento da oralidade. Em seguida, aponta que, por ser um teste aplicado em larga escala e na versão papel e lápis, a Matriz de Referência da PB considerou somente as habilidades dos eixos 1, 2 e 4, sendo que este último eixo não se caracteriza como um eixo separado, permeando toda elaboração do teste. Assim, a Matriz de Referência da PB foca na proficiência leitora e apresenta itens construídos para atender diretamente aos eixos 1) apropriação do sistema de escrita e 2) leitura, com o esclarecimento de que a compreensão e valorização da cultura escrita seria um eixo tratado de forma indireta na construção dos itens.

A definição da natureza do objeto educacional favorece a seleção do instrumento compatível com ela. Por exemplo, se o objeto educacional remete-se aos processos de escrita, a coleta deve incluí-la. A definição dos objetos educacionais tem o propósito de aclarar o que será considerado na elaboração das tarefas – os itens no caso de uma prova – propostas nos instrumentos, como ponto de partida, para, como ponto de chegada, garantir que a análise das respostas a essas tarefas seja de qualidade – característica – relativa a esse objeto, garantindo sua validade, e, com base em critérios, permitir a emissão de um julgamento de mérito dos resultados. Esse conjunto de elementos explicitados é o que institui a transparência do processo avaliativo e favorece, inclusive, sua crítica e meta-avaliação, dotando-o de maior objetividade.

O *levantamento e análise de informação relevante* sobre o objeto educacional é mais um componente fundamental da avaliação, expresso na definição de Lukas Mujika e Santiago Etxebarría (2009). Os autores ponderam que esse levantamento e a análise poderão ser de natureza qualitativa ou quantitativa, mas, para as duas situações,

indicam a importância de um cuidadoso e planejado processo, ao imprimir termos como: rigorosa, planificada, objetiva, fidedigna e válida, entre outros.

O teste da PB – Leitura é composto por itens de múltipla escolha que cobrem a matriz de avaliação e se caracterizam por abarcar conhecimentos e habilidades muito próximos daquilo que está sendo abordado em sala de aula, favorecendo o caráter formativo da avaliação e, simultaneamente, sendo mais inteligível aos(às) professores(as).

A definição dos contornos do objeto será fundamental também para a descrição e interpretação consciente e consistente dos dados, com o estabelecimento do critério de qualidade considerado satisfatório e que viabiliza a qualificação do objeto. A coleta de informações pode ocorrer de diferentes formas, devendo seguir sempre rigor metodológico. Contudo, as observações de performances em seminários e projetos bem como outras manifestações escritas e pictóricas, como os testes e portfólios, solicitadas no contexto da prática cotidiana, nem sempre recebem o rigoroso planejamento, quer quanto à definição do objeto educacional, quer quanto ao estabelecimento dos “critérios e referências” nos quais o juízo de valor estará baseado.

Se, de um lado, reconhecemos que nem todos os processos terão esse cuidado – por limitações materiais ou por não requererem –, visto que a avaliação informal ocorre com frequência na prática pedagógica, de outro, não se pode conceber que práticas avaliativas sejam marcadas, generalizadamente, pela ausência desses cuidados, com o risco de construir um juízo de valor em bases arbitrárias sobre informações coletadas impropriamente.

Como fator decisivo para a emissão de um julgamento, o critério de avaliação deve estar conectado aos processos de delimitação do objeto e da coleta de informações, sendo que essas etapas auxiliam a descrição da realidade observada e da atribuição de uma qualidade por meio de “sua comparação com um padrão de qualidade assumido como satisfatório” (LUCKESI, 2018, p. 151-152).

A definição ainda menciona a necessidade de “levantamento de informação de forma fidedigna e válida”. Essa característica se refere à precisão, que deve ser observada tanto nos testes como em outros instrumentos de coleta, e está vinculada ao criterioso processo de construção dos instrumentos e às respostas obtidas. Na medida

educacional, a precisão pode ser aferida e controlada por meio de análises Psicométricas, que serão discutidas no Capítulo 4.

O teste da PB – Leitura possibilitou a configuração de uma escala de referência quanto às tarefas cognitivas respondidas por crianças de todo o país. Essa escala de proficiência está dividida em cinco níveis, que contam, cada um, com uma descrição quanto aos domínios cognitivos dos(as) estudantes cujas proficiências aí se localizem, ou seja, uma interpretação pedagógica. A escala foi organizada a partir de um processo estatístico que, posteriormente, recebeu a interpretação de professores(as) alfabetizadores(as) e especialistas da área de alfabetização e letramento inicial. Foram esses profissionais que definiram qual seria o nível considerado satisfatório para estudantes do 2º ano do ensino fundamental, sendo indicados os níveis 4 ou 5.

No estabelecimento do critério para analisar a objeto da avaliação, é necessário buscar parâmetros objetivos, ou seja, deve-se pautar a qualidade em aspectos concretos projetados para o ato pedagógico, essencialmente um ato teleológico.

O teste oferecido pela PB refletiu uma prática avaliativa que leva em consideração os elementos da medida educacional que, por sua vez, contribuem para a definição não arbitrária de critérios de análise da realidade avaliada, qual seja a leitura, para auxiliar o juízo de valor. O juízo de valor é a comparação entre a realidade descrita, obtida mediante a coleta planejada de dados com o padrão de qualidade assumido como válido para o objeto em estudo. (LUCKESI, 2018, p. 54).

A definição de Lukas Mujika e Santiago Etxebarría (2009) encerra o processo com a tomada de decisões para otimizar o objeto educacional eleito inicialmente. Nesse ponto, anunciam a função central do processo, indiscutivelmente apontando para a forma de utilização dos resultados. Luckesi (2018), por sua vez, discorda que a tomada de decisão faça parte do ato de avaliar, embora considere-a fundamental para o gestor da ação pedagógica.

A decisão quanto ao nível satisfatório – ou adequado ou desejável – do resultado do objeto avaliado e a consequente decisão por uma ação pedagógica não podem ser relegados ao processo de medida. A decisão envolve ponderar as informações que a medida oferece com outros aspectos do contexto educacional, quais sejam, históricos-sociais, materiais, culturais, entre outros, para ensejar medidas pedagógicas que possam fazer com que cada estudante, estando abaixo do que era

esperado, possa avançar em sua aprendizagem no objeto em questão. Por exemplo, envolve considerar o nível de proficiência alcançado em relação ao que foi agregado à proficiência existente anteriormente; se o processo de ensino planejado, incluindo os recursos pedagógicos, foi de fato desenvolvido; se há um currículo individualizado a ser considerado na situação de deficiência etc.

O conjunto de documentos que compõem a PB – Leitura contém sugestões pedagógicas para cada um dos cinco níveis de proficiência, as quais também foram elaboradas por um grupo de profissionais e especialistas em alfabetização e letramento inicial. Essas sugestões, de maneira geral, caracterizam a etapa da tomada de decisão para otimizar o objeto.

Pelo exposto, pode-se perceber que medir não é avaliar, mas uma medida pode apresentar as características essenciais para a realização de um processo de avaliação, favorecendo seu caráter formativo.

Mesmo com o criterioso cuidado da PB – Leitura em definir o objeto de conhecimento, fornecendo teste que atenda a esse objeto e escala de proficiência interpretada, foram observados limites na versão papel e lápis da PB – Leitura, os quais serão determinantes para a proposta de uma versão informatizada e adaptativa para esse teste. Esses limites e os argumentos que justificam a construção do TAI da PB – Leitura serão discutidos na seção a seguir.

2.3 O teste na versão papel e lápis da PB – Leitura: características, potencialidades e limitações

O teste original da PB é um instrumento padronizado e disponibilizado para todo o território brasileiro na versão impressa pelo Inep, com o objetivo de auxiliar o diagnóstico da alfabetização e do letramento nas séries iniciais do ensino fundamental. Sua criação procurou atender a uma necessidade no âmbito das políticas públicas de identificar rapidamente os problemas na aprendizagem e permitir intervenções que promovessem avanços em dificuldades que a educação brasileira tem enfrentado, sobretudo no processo de alfabetização e letramento inicial. Conforme exposto no próprio portal do Inep:

> A Provinha Brasil é um instrumento de avaliação processual que apoia o trabalho cotidiano dos professores. Ela é composta de cinco instrumentos, todos eles disponibilizados no portal: teste de leitura/aluno, teste de leitura/aplicador, teste de matemática/aluno, teste de matemática/aplicador, guia de Correção. Cada prova contém 20 questões. O material do professor tem comentários sobre o que cada questão é capaz de avaliar, permitindo que o exame se torne um instrumento de intervenção pedagógica. (PORTAL INEP[13])

Nas séries históricas dos resultados de avaliações externas no Brasil, notadamente do Sistema de Avaliação da Educação Básica (Saeb), mais da metade dos estudantes do 5º ano do ensino fundamental apresentam proficiências em testes que avaliam a leitura na Prova Brasil[14] consideradas baixas e cerca de 22,2% dos(as) alunos(as) brasileiros(as), em 2014, e 21,7%, em 2016, apresentavam patamares considerados insuficientes de domínio da leitura na Avaliação Nacional da Alfabetização (ANA)[15]. As expressivas taxas de reprovação nos anos iniciais do ensino fundamental são igualmente reveladoras dessas dificuldades, para as quais o que mais pesa nos

---

[13] Disponível em:<http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/provinha-brasil-ja-esta-disponivel-no-portal-inep/21206>. Acesso em 14 jan. 2018.

[14] Avaliação envolvendo os alunos do 5º ano e do 9º ano do ensino fundamental das escolas públicas brasileiras com o objetivo principal de avaliar os níveis de proficiência em Língua Portuguesa, com foco em leitura, e Matemática, com foco em resolução de problemas. A Prova Brasil foi incorporada ao Sistema de Avaliação da Educação Básica (Saeb) pela Portaria nº 931 (BRASIL, 2005).

[15] Avaliação que envolveu alunos do 3º ano do Ensino Fundamental das escolas públicas brasileiras com o objetivo de avaliar os níveis de alfabetização e letramento inicial em Língua Portuguesa, alfabetização em Matemática e condições de oferta do Ciclo de Alfabetização das redes públicas. A ANA foi incorporada ao Saeb pela Portaria nº 482 (BRASIL, 2013b).

critérios dos professores(as) é o domínio dos(as) estudantes em relação ao sistema de escrita alfabética e letramento inicial.

Não se ignora que esses resultados estão fortemente associados às situações de vulnerabilidade dos estudantes, mas também é certo que fatores internos às escolas, relativos aos seus procedimentos pedagógicos, exercem influência na matéria (ALAVARSE, 2009; CRAHAY, 2002). Dentre os fatores internos às escolas e relevantes para os resultados, estão as práticas avaliativas, não poucas vezes realizadas sem as devidas validade e fidedignidade (NEVO, 1998). Sob esse aspecto, constata-se a inexistência de dispositivos que diagnostiquem níveis de proficiência dos(as) alunos(as) em leitura, de maneira rápida e precisa, eventualmente em grande escala, e o teste em papel e lápis da PB foi uma tentativa de fornecer instrumentos nessa perspectiva, embora o processo manual de contagem de acertos para obtenção dos resultados ainda exija tempo de dedicação das equipes educacionais.

No processo da alfabetização, conforme aponta Soares (2016), a escrita e as práticas de leitura são consideradas produtos culturais dos mais notáveis, tornando o complexo processo de alfabetização e letramento inicial ainda mais importante do ponto de vista social e político.

Mesmo considerando as polêmicas sobre suas abordagens pedagógicas (Cf. MICOTTI, 2013), sua avaliação (Cf. ESTEBAN, 2009; MORAIS, 2012) e as dificuldades existentes na delimitação da leitura, enquanto objeto de avaliação, especialmente pela diversidade cultural do Brasil, a PB – Leitura possui características de um processo de avaliação com validade nacional, mais um aspecto que foi decisivo para a seleção desse teste.

Além da importância político-pedagógica da PB – Leitura e dos elementos necessários para a constituição de uma avaliação educacional, apontados no capítulo 2, outras características do teste impresso foram consideradas para a construção do TAI.

Os testes em papel e lápis da PB – Leitura eram disponibilizados no sítio do Inep e distribuídos nacionalmente em dois períodos distintos do ano letivo, com indicação para ser aplicado no segundo ano do ensino fundamental. No início do ano letivo, aplicava-se o teste 1 e, no final do ano letivo, o teste 2. A distribuição nacional dos dois testes ocorreu de 2008 a 2015. Em 2016, somente o teste 1 foi impresso e

enviado nacionalmente; já o teste 2 foi disponibilizado apenas no sítio do Inep[16]. Cada teste era composto de um caderno com 20 itens para Língua Portuguesa, com foco na leitura, e outro caderno com 20 itens para Matemática, com foco na resolução de problemas. Optou-se por não transpor os itens de Matemática na construção do TAI pelo papel relevante da leitura nos anos iniciais da escolarização.

Por parte do Inep e das secretarias de educação que aderiam à aplicação da PB, a complexidade estava em distribuir os *kits* contendo os cadernos de itens para os estudantes e os guias de aplicação e de interpretação dos resultados para os professores(as) e gestores(as). Por parte das unidades educacionais, a dificuldade estava na distribuição dos *kits* para os(as) professores(as) e na orientação aos docentes, envolvendo o planejamento da aplicação em cada turma, a obtenção do número de acertos por aluno(a), a alocação deles(as) em cada um dos cinco níveis identificados no guia de interpretação dos resultados e a análise das intervenções necessárias para aqueles que não se encontravam nos níveis considerados adequados.

A proposição de dois testes durante o ano letivo tinha o propósito de atender à natureza diagnóstica da avaliação, ou seja, para acompanhar se as intervenções pedagógicas implementadas após o teste 1 tinham efeitos nas proficiências aferidas no teste 2. Os testes da PB – Leitura eram formulados para permitir que professores(as) e gestores(as) identificassem, com base em instrumentos padronizados, os avanços e as dificuldades na alfabetização e no letramento de seus estudantes. As intervenções e análises pedagógicas eram subsidiadas pelos documentos (guias) do *kit* destinados aos professores(as) e gestores(as).

As opiniões se dividem quanto à identificação do caráter somativo e externo ou formativo e interno da PB. De modo geral, mesmo sendo um instrumento padronizado, elaborado por uma equipe central e externa à unidade escolar, a PB mantinha alguns aspectos de avaliação interna e formativa, uma vez que o(a) próprio(a) professor(a) da turma se encarregava da aplicação, da análise e da interpretação dos resultados e

[16] Em oito de agosto de 2016, a Provinha Brasil passou a ser disponibilizada apenas no sítio do Inep e os responsáveis por ela apresentaram como justificativa: “as restrições financeiras e as alterações curriculares que envolviam a Base Nacional Comum Curricular” (TOKARNIA, 2018).

o teste aferia o domínio de habilidades muito próximas do trabalho pedagógico, alinhadas com a prática da sala de aula.

Ao transpor o teste na versão papel e lápis para o TAI, haveria condições de permitir que aplicações fossem realizadas em diferentes momentos do ano escolar, para um(a) aluno(a) ou para toda a turma, fornecendo o resultado imediatamente após o término da administração, tornando as intervenções mais ágeis.

Os testes impressos eram compostos por 20 questões, mais formalmente itens de múltipla escolha, também denominados objetivos[17], e o(a) próprio(a) professor(a) da turma era o aplicador da prova, recebendo orientações muito semelhantes às realizadas em uma avaliação em larga escala, dadas no guia de aplicação que fazia parte do *kit*. Na edição de 2016, (BRASIL, 2016a) o *kit* da PB – Leitura era composto de:

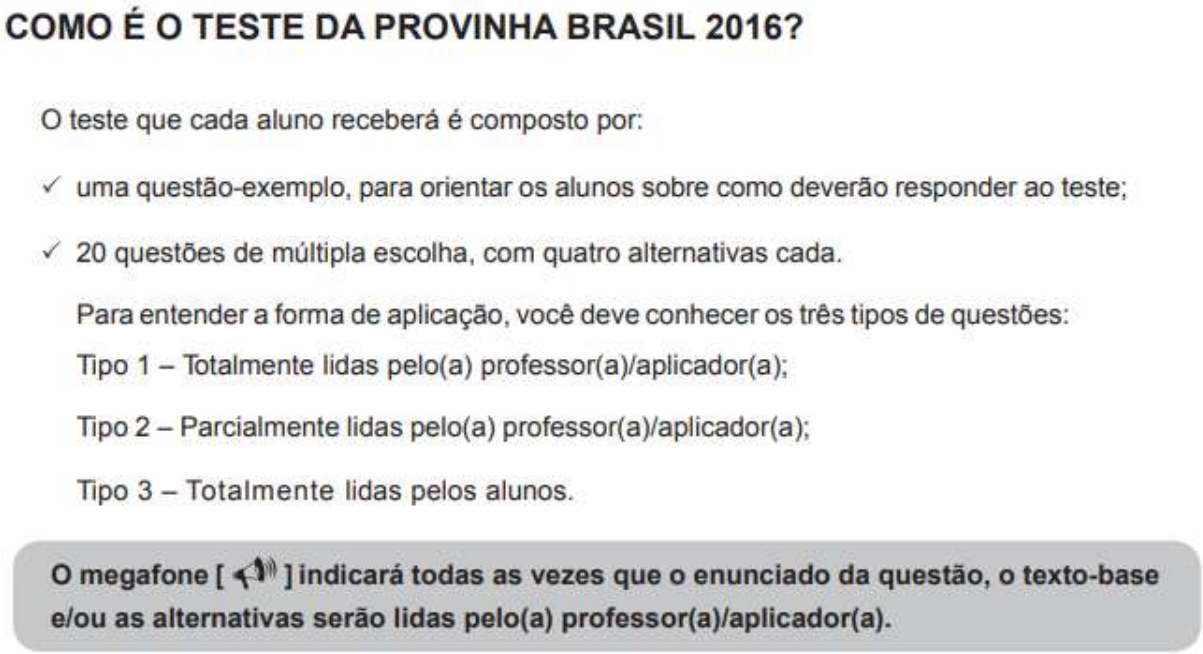
a) **caderno do aluno**, contém o teste em leitura com 20 itens, tipicamente administrado em papel e lápis, os quais são apresentados sem os trechos que serão lidos pelo(a) professor(a);
b) **guia de aplicação**, contém orientações sobre a aplicação do teste, abrangendo os itens na integralidade e explicações sobre a condução da leitura em voz alta no teste;
c) **guia de correção e interpretação**, um documento com orientações sobre a construção do teste, contendo: os aspectos sobre a concepção do teste, a indicação das respostas certas (gabaritos) dos itens, a sugestão de um quadro para registro dos acertos da turma, um quadro com a correspondência entre acertos e os níveis da escala de proficiência. Também constam desse documento a interpretação pedagógica de cada nível da escala e as sugestões de intervenção.

Um dos desafios da aplicação do teste em papel e lápis está na padronização da leitura em voz alta de parte de alguns itens, realizada pelo professor(a)/aplicador(a), conforme mostra a Figura 1. Essa leitura, embora necessária para atender alunos(as)

[17] É denominado objetivo somente o item cujo gabarito é considerado por um grupo de especialista como a única resposta correta.

que ainda se encontravam sem autonomia leitora, interferia profundamente na confiabilidade dos resultados, por comprometer a padronização das condições de aplicação do teste. Outro entrave estava na necessidade de coordenar o avanço simultâneo dos alunos em cada uma das 20 questões, para que fosse possível o acompanhamento da proficiência em leitura por todos os alunos da sala, o que contribuía para a dispersão no momento da aplicação.

Figura 1 – Forma de aplicação da PB – Leitura, por tipo de questão



Fonte: Brasil (2016a).

Essa dispersão ocorria principalmente porque o conjunto de itens é o mesmo para todos(as) os(as) alunos(as), independentemente da heterogeneidade de suas proficiências. Para exemplificar esse problema da aplicação, um(a) aluno(a) com proficiência do nível 1, que se situa em torno de cinco acertos das 20 questões do teste, devia continuar a responder aos demais itens, cuja dificuldade aumentava gradativamente, ficando numa situação extremamente desconfortável e, por vezes, constrangedora. No outro extremo, um(a) aluno(a) com proficiência no nível 5, que apresenta total de acertos entre 18 e 20 itens, podia ser desmotivado ao ter que responder a itens muito fáceis, do início da prova. Essa dinâmica de aplicação, necessária na aplicação da versão impressa do teste, impossibilitava que cada aluno(a) avançasse ao seu ritmo nos itens do caderno de prova, pois são ocultadas

as partes do item que serão lidas pelo professor(a)/aplicador(a) no Caderno do aluno, inviabilizando a resolução autônoma do teste, mesmo para crianças mais proficientes.

Ao final da aplicação da versão impressa, os cadernos de itens de cada aluno(a) deviam ser recolhidos e suas respostas, cotejadas com as respostas certas – os gabaritos indicados pelo Inep –, para que fosse assinalado se houve acerto ou erro em cada item. Finalmente, os acertos deviam ser somados e o total contrastado com um guia de acertos para localizar o nível – de 1 a 5 – de cada respondente, decorrendo daí a interpretação pedagógica.

Subjacente a esse processo de estimativa da proficiência dos estudantes encontra-se a opção, por parte do Inep, de utilização do modelo de um parâmetro da TRI na parametrização dos itens da PB – Leitura e com isso as proficiências dos(as) alunos(as) passa a ser consequência direta da contagem dos acertos observados no teste. Os guias de interpretação dos resultados, emitidos pelo Inep, exprimem os níveis de proficiência da escala em uma tabela de acertos, possibilitando a alocação do(a) aluno(a) em um dos 5 níveis de proficiências sempre na mesma escala e com base no número de acertos. Cada conjunto de itens permite alocar os(as) estudantes numa escala dividida em cinco níveis, numerados do 1 ao 5, sendo que o nível 1 refere-se ao menor domínio aferido para a proficiência em leitura e o 5, ao maior. Mantida a mesma escala para aferição da proficiência, a interpretação pedagógica realizada com base na ancoragem dos itens permite que sejam identificados os avanços e as dificuldades, bem como as possibilidades de intervenção.

O processo envolvido na contagem de acertos e alocação dos(as) estudantes nos níveis de proficiência leitora para constituição dos resultados, somados às dificuldades de aplicação relatadas anteriormente, demandavam expedientes que tornavam a PB – Leitura um recurso que, apesar de seu potencial pedagógico, exigia tempo para o planejamento da aplicação, para a conferência dos acertos e, sobretudo, para a tarefa de análise dos resultados.

Os problemas advindos da aplicação, da tabulação e da padronização da leitura dos itens, em grande parte, seriam equacionados pelo teste eletrônico (TBC) e TAI da PB – Leitura.

Outra vantagem em transpor o teste da PB – Leitura para um TAI está no banco de itens necessário para um bom funcionamento da ferramenta. Qualquer iniciativa

exigiria considerável investimento inicial na obtenção de itens válidos e calibrados. Na PB – Leitura, ao contrário, seria possível partir de cerca de 320 itens, disponibilizados pelo Inep ao longo de suas oito edições realizadas até o momento, todos parametrizados pela TRI.

O Projeto TAI da PB – Leitura proporcionou documentos formulados por especialistas e assessores que desenvolveram os testes da PB – Leitura (BRASIL, [2011b], 2012c, 2013a, 2014). Esses documentos continham: a definição do objetivo da avaliação; a habilidade atendida em cada item; o modelo de análise dos itens, dado pela Teoria Clássica de Testes (TCT) e Teoria da Resposta ao Item (TRI), o modelo de um parâmetro; os aspectos que garantem a idoneidade do traço latente medido; os parâmetros de 40 itens da edição de 2015 da PB – Leitura; a definição da escala de proficiência em leitura e dos cortes que determinam os níveis de proficiência.

Os parâmetros dos itens e a descrição da escala de proficiência nos pressupostos da TRI, contidos nesses relatórios, foram essenciais para a transposição do teste em papel e lápis para o TBC e posteriormente para o TAI.

Em linhas gerais, essas orientações indicavam os fundamentos da avaliação educacional realizada na PB – Leitura, com fulcro na teoria do traço latente. A medida do traço latente centra nos aspectos observáveis do comportamento humano para inferir as características que não são diretamente observáveis, a exemplo da aprendizagem que será inferida por meio das respostas (desempenho) aos itens. Nesse caminho, a medida contou com o auxílio de metodologias psicométricas, como a Teoria Clássica dos Testes (TCT) e Teoria da Resposta ao Item (TRI) (Cf. BORSBOOM, 2003, 2005; DE AYALA, 2009; HAMBLETON; SWAMINATHAN, 1985; LORD; NOVICK, 1968; PASQUALI, 2007, 2013) na análise dos resultados observáveis.

A medição de um domínio ou de uma proficiência, amparada na medida do traço latente, destaca o problema do erro associado, em maior ou menor grau, a toda e qualquer medida. Esse erro, conforme aponta a literatura (ALAVARSE; MELO, 2013a, 2013b; HAMBLETON; JONES; ROGERS, 1993; KLEIN, 2013; OLEA; PONSODA, 2003; OLEA; PONSODA; PRIETO, 1999; PASQUALI, 2013) e segundo foi investigado para o teste da PB – Leitura, estaria presente em maior grau para estudantes com proficiências nas extremidades da escala, por terem sido submetidos a um instrumento com sequência fixa de itens, como ocorre na versão impressa do teste.

Esse erro é minimizado quando os itens do teste apresentam graus de dificuldade semelhantes à proficiência do(a) respondente.

No entanto, há outra vantagem que a TRI oferece para um TAI. Com ela, tem-se o pressuposto da invariância dos parâmetros dos itens (COUTO; PRIMI, 2011), que possibilita ao(à) respondente, mesmo recebendo uma sequência diferente de itens (testes diferentes para cada respondente), ter sua proficiência estimada na mesma escala, que pode ser apresentada com uma interpretação pedagógica, como ocorre com a escala de proficiência da PB, características abordadas no capítulo seguinte.

3 O TESTE ADAPTATIVO INFORMATIZADO (TAI)

Os testes utilizados nas avaliações externas no Brasil, embora construídos segundo normas internacionais para elaboração de avaliações em testes de larga escala e com utilização de sofisticados modelos da TRI para análise dos itens e respostas, ainda apresentam, conforme aponta a literatura, os limites típicos dos instrumentos de avaliação administrados com o uso de papel e lápis.

Um desses limites refere-se ao número de itens ao qual a população de respondentes é submetida. Os testes lineares, nos quais o conjunto de itens é fixo e idêntico para o grupo de respondentes, devem apresentar itens com níveis de dificuldade que possam atender à variabilidade nas proficiências da população; e em cada nível de dificuldade, o número de itens deve ser razoável para garantir que a aferição da proficiência da subpopulação de correspondente domínio possa ocorrer com erro de medida em patamares aceitáveis, ou seja, ter boa precisão nos resultados da maioria dos(as) respondentes (YAN, VON DAVIER; LEWIS, 2014). Como consequência, o teste fica extenso. Decorrente desse conjunto fixo de itens, o(a) respondente de proficiência mais baixa é submetido(a) tanto às questões com nível de dificuldade próximo ao que consegue responder quanto às questões com nível de dificuldade maior, mais adequadas aos(às) respondentes com proficiências mais altas e vice-versa. Essa condição, além de trazer inconvenientes para o(a) respondente, que no caso citado pode se sentir constrangido(a) por se considerar inapto(a) diante de questões difíceis, acarreta maior erro de medida na aferição da proficiência, conforme será abordado mais adiante.

Outra limitação observada no teste de papel e lápis é a falta de rapidez e praticidade na obtenção dos resultados. Esses testes exigem tempo para o processamento dos acertos, a estimação da proficiência de cada respondente e a elaboração de uma escala de medida interpretada nos termos do objeto da avaliação. Igualmente, nos testes em papel e lápis estão ausentes os itens em formato multimídia, que possibilitam a incorporação da tecnologia e de objetos midiáticos, potencializando a coleta de informações sobre conhecimentos e competências que não poderiam ser diagnosticados em itens na versão impressa.

O crescente desenvolvimento da tecnologia também impactou de forma significativa o cenário da avaliação educacional (Cf. FOSTER, 2015; LUECHT, 2013; MAGIS; MAHALINGAM, 2015). Além dos incrementos nas análises computacionais, que possibilitaram a produção de *softwares* para análises estatísticas sofisticadas, como da Teoria da Resposta ao Item (TRI), os testes puderam ser administrados por meio de computadores (*desktops*) ou dispositivos portáteis como *laptops, tablets* e celulares. Wang e Shin (2010) ressaltaram que, ao entregar um teste de papel e lápis, convencional, por meio de um computador, dois tipos de informatização podem ser implementados. O primeiro tipo de informatização para o teste confere a ele a categoria de Teste Baseado em Computador (TBC)[18] e contempla a mudança da versão impressa para o meio digital, na qual os itens são transpostos para a tela de um computador ou outro dispositivo portátil, continuando idênticos ao formato anterior. Além disso, os métodos e procedimentos de pontuação (aferição da proficiência) do TBC podem continuar os mesmos, visto que a sequência e quantidade de itens continuam fixas, como na versão impressa e por esse motivo o TBC é considerado um teste linear. Quanto ao termo linear, adotado na literatura, será apresentado um esclarecimento na subseção 3.2. A mudança da versão impressa para TBC, portanto, pode envolver apenas uma alteração no modo de administração. O outro tipo de informatização do teste atribui a ele o caráter de Teste Adaptativo Computadorizado (TAC)[19] ou Teste Adaptativo Informatizado (TAI), no qual não apenas o meio de administração muda (do papel para o digital), mas também o algoritmo de entrega de testes altera o TBC, transformando-o de linear para adaptativo. A abordagem adaptativa do teste coloca a tecnologia para interferir no modo de selecionar os itens que farão parte do teste individualizadamente, por interagir com as respostas dadas pelo(a) respondente. A tecnologia ainda interfere no modo de aferir a proficiência, pois serão realizadas aferições parciais para ajuste da sequência de itens (ou grupo de itens no caso de testes de múltiplos estágios) de modo que sejam adaptados à capacidade de cada participante do teste. Portanto, tanto o modo de administração quanto a abordagem adaptativa diferenciam o TAI de um teste homólogo em papel e lápis.

---

[18] Do inglês *Computer-Based Tests* (CBT).
[19] Do inglês *Computerized Adaptive Testing* (CAT).

Essas reflexões vão ao encontro da categorização apresentada por Yan, Lewis e von Davier (2014), de modo que os TBC podem ser divididos em lineares ou adaptativos. Os TBC lineares são versões eletrônicas dos testes em papel e lápis, mantendo o mesmo número e a mesma sequência de itens para todos os(as) respondentes. Os TBC adaptativos, além de serem administrados em ambiente informatizado, apresentam um algoritmo preparado para oferecer um teste individualizado e diferente para cada respondente, sendo que os itens que farão parte desse teste individualizado são selecionados para atender melhor a proficiência de cada respondente. Os TBC adaptativos, são identificados por testes adaptativos informatizados (TAI).

Nas próximas subseções, serão exibidas as principais características, vantagens e limitações do TBC tanto na versão linear como na versão adaptativa.

3.1 Os testes baseados em computadores (TBC): caracterizações e limitações

Os Testes Baseados em Computadores (TBC) são os testes indicados para contornar parte das limitações dos testes em papel e lápis. Esses testes incorporam as tecnologias de informação e comunicação (TIC) ao apresentarem as questões por meio digital.

Nesta seção, são apresentadas as características dos TBC lineares, cuja primeira vantagem, consiste em proporcionar a imediata obtenção de resultados do teste. Diferente da administração por meio de papel e lápis, que requer uma série de etapas até o processamento dos resultados, envolvendo a tabulação dos acertos e análise estatística para aferição das proficiências, os TBC proporcionam a obtenção automatizada de resultados. Duas fundamentais etapas são eliminadas: a) o transporte das respostas para uma folha de respostas, aspecto que adquire especial importância para respondentes de determinadas faixas etárias, como crianças e idosos, ou inexperientes na realização de testes; e b) a digitação ou digitalização das respostas para constituição do banco de dados para análise estatística. Nos TBC a interação dos(as) respondentes com o dispositivo eletrônico usado para sua aplicação – um computador ou um *tablet* – permite a constituição do banco de dados das

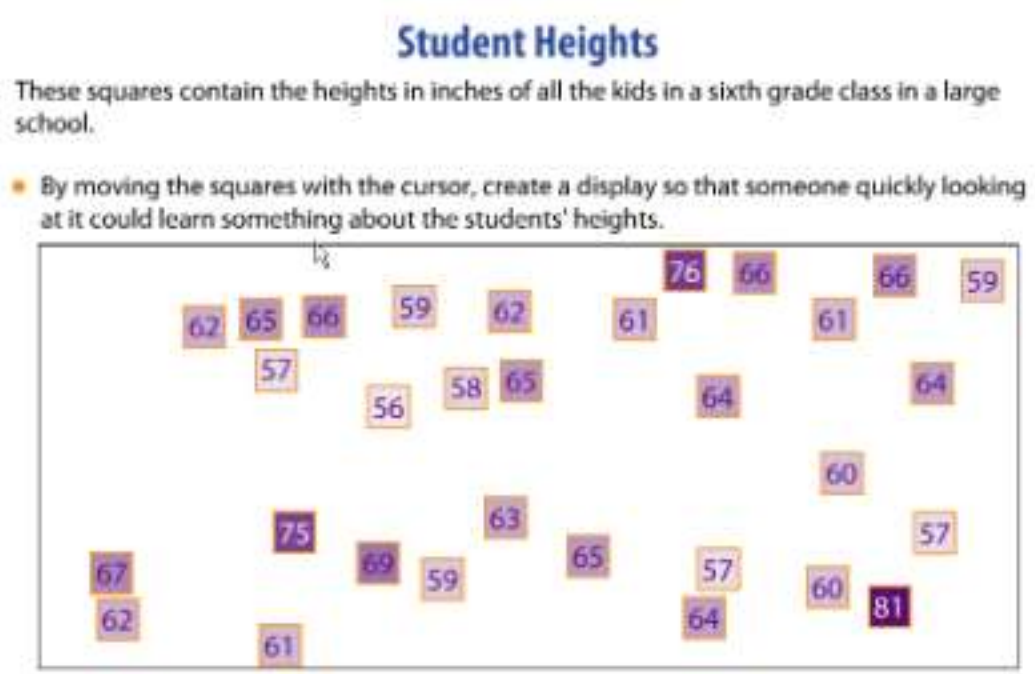
respostas, com rapidez na obtenção dos resultados, especialmente em testes em larga escala, e análises estatísticas automatizadas.

A segunda vantagem consiste em permitir incorporar à elaboração dos itens o uso de ferramentas tecnológicas, diversificando e ampliando as tarefas ou problemas propostos aos(às) respondentes. Os novos itens, construídos para o TBC, podem lançar mão das ferramentas tecnológicas de duas diferentes maneiras, possibilitando a ampliação: a) dos modos de apresentar os contextos e/ou objetos auxiliares/suportes na reflexão proposta pelo item e que mobilizam uma resposta do(a) respondente, podendo integrar movimento e som a figuras, gráficos, textos e desenhos já utilizados nos itens de testes de papel e lápis; e b) das operações cognitivas solicitadas, suplantando as possibilidades de expressar escolhas, descrições, identificações, comparações, relacionamentos, análise e avaliações em torno dos fatos e fenômenos, ou linguagens, objetos da aferição pretendida. Os recursos e dispositivos eletrônicos, com interfaces gráficas e acesso a objetos educacionais, apontam para uma nova modalidade de itens e a proposição de atividades inteiramente novas, impossíveis nos itens elaborados para a prova de papel e lápis. Nessa perspectiva, Scalise (2009) apresenta novos tipos de itens que agregam texto, imagem, som e animação gráfica e, por esse motivo, permitirão ao(à) respondente a construção de respostas que são automaticamente aferidas, indicando que novos tipos de itens poderão incluir desde respostas restritas, convencionalmente identificadas como questões de múltipla escolha, como as parcial e totalmente construídas.

Como exemplo, a Figura 2 mostra um item em que os(as) respondentes devem organizar os quadrados que contêm os números representativos das alturas de alunos(as) de uma classe. Os quadrados devem ser arrastados e organizados em uma espécie de gráfico, facilitando a análise das informações, por exemplo, saber qual é a menor, qual é a maior, qual é a mais frequente. A construção desse gráfico, que não é rígida, podendo ser em barras verticais ou horizontais, permite mais de uma única possibilidade de acerto e pode ser corrigida de maneira automatizada, a partir de programação que apresente as diferentes soluções previstas como corretas. Nos testes de papel e lápis, esse item seria de resposta construída, cuja correção teria implicações no custo, no tempo e na padronização das interpretações das respostas.

Em uma avaliação em larga escala em papel e lápis, essas implicações costumam inviabilizar a avaliação desse tipo de habilidade.

Figura 2 – Exemplo de item inovador, de resposta construída com figuras



Fonte: Projeto desenvolvido por Scalise[20].

Um TBC admite vantagens em relação ao teste de papel e lápis, entretanto o principal obstáculo continua sendo o fato de o conjunto de itens do teste ser fixo e idêntico para todos os(as) respondentes, pois já está predeterminado por aqueles que projetam o teste, não importando a proficiência do(a) respondente nem o tempo de resposta. A predeterminação dos itens é uma das mais significativas fontes de erro de medida, pois pode ocorrer que o(a) respondente seja submetido(a) a itens que estejam muito acima ou muito abaixo da sua proficiência (ALAVARSE; MELO, 2013a, 2013b; KLEIN, 2013). Em consequência, o(a) respondente recebe um teste que está muito distante de suas condições, podendo desistir ou não de responder-lhe com seriedade e o esforço em responder-lhe não atribuirá maior fidedignidade à estimação de sua

[20] Página do Projeto disponível em: <https://pages.uoregon.edu/kscalise/taxonomy/ taxonomy.html>. Acesso em: 13 jan. 2018.

proficiência. Para contornar essa limitação tem-se o TAI que é um TBC com características adicionais.

3.2 Os testes adaptativos informatizados (TAI): apresentação e características

Como citado anteriormente, os TBC podem ser lineares ou adaptativos. Os TBC adaptativos recebem a denominação TAI e se dividem em testes adaptativos: para cada item ou para cada bloco de itens, neste último caso são denominados de testes de múltiplos estágios (Cf. OLEA; PONSODA; PRIETO, 1999; YAN; VON DAVIER; LEWIS, 2014).

O presente estudo abordará os TAI para cada item. Os testes adaptativos informatizados (TAI), na literatura também identificados por testes adaptativos computadorizados (TAC), trouxeram inúmeras vantagens para os testes conforme apontam Alavarse e Catalani (2016a, 2016b), Alavarse e Melo (2013a, 2013b), Alavarse et al. (2017; 2018a), Barrada (2012), Chang (2012), Cheng (2009), Davey e Pitoniak (2006), Klein (2013), López-Cuadrado, Pérez e Armendariz (2005), Ponsoda Gil et al. (2004), Way et al. (2015), entre outros. As vantagens identificadas na revisão da literatura apontam que os TAI têm a possibilidade de:

a) integrar as ferramentas tecnológicas na elaboração das questões (itens) da prova, apresentando novos formatos para os itens;
b) aferir novos tipos de habilidades;
c) administrar um teste a pequenos grupos de estudantes, aferindo a proficiência em uma única escala;
d) fornecer a pontuação imediatamente após o término da aplicação;
e) administrar um teste em intervalos de tempo mais frequentes (o que é chamado de teste contínuo) e flexibilizar o agendamento do teste;
f) analisar os resultados com maior rapidez;
g) fornecer diagnóstico para pais, mães ou responsáveis, professores(as), alunos(as) e gestores(as) da escola, modulando a linguagem e as informações pedagógicas para cada público-alvo;

h) registrar informações longitudinais para monitorar o desenvolvimento dos(as) estudantes;
i) personalizar o teste de acordo com os conhecimentos do(a) respondente;
j) reduzir a quantidade de itens (comprimento do teste) e o tempo de aplicação do teste;
k) conferir maior precisão às estimativas da proficiência, permitindo que o teste seja mais informativo tanto do ponto de vista psicométrico, com o erro de medida menor e mais homogêneo em todo o intervalo de variação da proficiência estimada, como do ponto de vista pedagógico, com a cobertura das especificações de conteúdo dadas na matriz de avaliação;
l) melhorar a segurança dos itens do teste, ao apresentar diferentes conjuntos de itens aos(às) respondentes, a depender das características do banco de itens existente;
m) fornecer, imediatamente após o término da aplicação, informações de diagnóstico para um domínio específico e relatórios, diretamente relacionados com os processos de ensino e aprendizagem da sala de aula, possibilitando informações essenciais para redirecionar o trabalho do(a) professor(a) ou para indicar intervenção individualizada para determinados(as) alunos(as).

É importante dizer que algumas das vantagens elencadas não são específicas do TAI e podem também estar relacionadas ao TBC, como as expressas nos tópicos de (a) a (h), mas a capacidade de ajustar o conjunto de itens do teste às características cognitivas dos(as) respondentes e fornecer informações de diagnóstico para um domínio específico e para identificação individualizada de intervenções educacionais é um desafio muito maior e que só o TAI pode enfrentar. Chang (2012) destacou que a maioria dos TAI foi desenvolvida para exames de admissão de larga escala de alta exigência, nos quais estimar com precisão a pontuação total verdadeira é a principal preocupação para o *design* de algoritmos de seleção de itens. Contudo, o interesse dos(as) professores(as) pode ir além da pontuação total e se voltar para o interesse em receber *feedback* para a intervenção pedagógica específica de seus (suas) estudantes, ou *feedback* para os(as) alunos(as) sobre as necessidades de estudos individualizados, potencializando a utilidade do teste.

Para que um TBC se torne um TAI acrescenta-se à administração informatizada dos itens um algoritmo de seleção de itens (Cf. OLEA; ABAD; BARRADA, 2010; YAN;

VON DAVIER; LEWIS, 2014). O objetivo central desse algoritmo é a constituição de testes particularizados, conforme a proficiência dos(as) respondentes. Para seleção dos itens, consideram-se as respostas dadas aos primeiros itens e a estimativa parcial da proficiência realizada com base neles. O item a ser ofertado em seguida será escolhido no banco de itens com o propósito de tornar as próximas estimativas da proficiência mais precisas, até que seja atendido o critério de parada. Como uma nova estimativa parcial é realizada para cada item respondido, o teste é caracterizado como um teste adaptativo ao nível do item.

Quanto ao termo linear usado na literatura para diferenciar o TBC que não é adaptativo, denota-se que não é suficiente para diferenciar o teste aplicado em meio digital, que mantém características dos testes em papel e lápis, de um TAI. As diferenças são mais complexas e, embora tenham sido cunhadas as expressões como “testes lineares”, “testes fixos” e “testes convencionais”, percebem-se diferenças que merecem melhores esclarecimentos. Para exemplificar essa complexidade, considere-se um teste planejado para ter número fixo e idêntico de itens para todos(as) os(as) respondentes, quando administrado em papel e lápis, ainda que o teste seja planejado para ser respondido em sequência linear e com o mesmo número de itens para todos(as), o(a) respondente pode optar por responder a essa sequência de itens não linearmente, retornando e saltando itens ou deixando sem respostas outros, conforme sua decisão. Um número *n* fixo e igual de item pode ser apresentado para todos(as) respondentes, a exemplo da PB e do Enem[21].

No Enem, é comum que os(as) respondentes saltem as questões mais difíceis e retornem a elas depois de resolverem as consideradas de fácil e média dificuldade, podendo inclusive retornar para modificar respostas já dadas, ou deixar de responder itens, adotando a resolução em uma sequência não linear e responder a um número não fixo de itens. Na PB, pela forma de condução realizada pelo professor(a)/aplicador(a), que faz a leitura parcial ou total de questão por questão, o(a) respondente geralmente preserva a sequência linear de respostas, embora o

[21] Exame aplicado pelo Governo Federal do Brasil com o objetivo de ser uma avaliação de desempenho de estudantes de escolas públicas e particulares do ensino médio. Em 2009, tornou-se também uma avaliação que seleciona estudantes de todo o país para instituições federais de ensino superior e para programas de financiamentos do Governo Federal.

número fixo de itens não seja mantido, pois alunos(as) podem não responder a certos itens.

De outra maneira, o teste em meio digital, seja adaptativo ou não, pode ser configurado para que se responda aos itens em uma dada sequência, sem que seja possível saltar ou retornar a itens já solucionados. Esse exemplo revela duas perspectivas na categorização dos testes: uma delas, do sujeito que planeja o teste, o(a) avaliador(a); e a outra, do sujeito que responde o teste, o(a) avaliado(a). Outro exemplo pode ser o Saeb, no qual o número de itens planejado é fixo para cada respondente da turma, embora eles(as) sejam submetidos(as) a testes com itens diferentes, devido à organização por Blocos Incompletos Balanceados (BIB). Os testes são diferentes, mas não personalizados quanto ao domínio do(a) respondente como no TAI. Nos cadernos de prova do Saeb, planejados com uma sequência fixa de itens, devido ao modo de aplicação, no qual o aplicador conduz a resolução bloco por bloco, pode haver uma certa linearização nas respostas. Em relação ao número de itens, embora planejados para terem o mesmo número de itens no caderno, como há a possibilidade de deixar itens sem respostas, a quantidade de itens respondida não seria fixa.

Assim, a característica de um teste, em papel e lápis ou em meio digital, é tratada segundo o planejamento do(a) avaliador(a), mas é necessário considerar também a forma como esse teste pode ser respondido. Nos testes informatizados, sejam do tipo TBC ou TAI, esses aspectos podem ser controlados na programação. A configuração pode ou não permitir saltar itens, deixando-os sem resposta; retornar a itens já respondidos para modificar respostas, fixar o comprimento do teste. Esses aspectos combinados estariam ou não garantindo a característica de sequência linear e a quantidade fixa de itens para o teste.

O quadro a seguir mostra uma síntese das diferentes possibilidades.

Quadro 1 – Características dos testes em lápis e papel, TBC e TAI

	Características dos testes	
	Quanto ao número de itens	**Quanto à sequência de itens**
Papel e lápis sem BIB	Fixo para o(a) avaliador(a) e variável para o(a) respondente	É linear para o(a) avaliador(a) e não linear para o(a) respondente
Papel e lápis com BIB	Fixo para o(a) avaliador(a) e variável para o(a) respondente	É linear para o(a) avaliador(a) e para o(a) respondente, mas pode se tornar não linear para o(a) respondente
TBC	Fixo para o(a) avaliador(a) e para o(a) respondente, mas pode se tornar variável para o(a) respondente	É linear para o(a) avaliador(a) e para o(a) respondente, mas pode se tornar não linear para o(a) respondente
TAI	Fixo para o(a) avaliador(a) e variável para o(a) respondente, mas pode se tornar variável para o(a) avaliador(a) e para o(a) respondente	É linear para o(a) avaliador(a) e para o(a) respondente, mas pode se tornar não linear para o(a) avaliador(a) e para o(a) respondente

Fonte: a autora.

Conforme ressalta Piton Gonçalves (2004), um TAI sugere uma maneira diferenciada de avaliar, visto que seleciona as questões conforme os níveis de proficiência estimados provisoriamente para o(a) respondente, atribuindo sentido ao termo adaptativo na denominação do teste. O algoritmo lança a primeira questão e, a partir da performance do(a) respondente, faz uso da Teoria de Resposta ao Item (TRI) para estimar a proficiência, e tanto as questões selecionadas como a quantidade delas (comprimento do teste) serão distintas e adequadas ao conhecimento de cada respondente. Neste caso, considerando que os(as) estudantes tendem a receber questões mais adequadas aos seus conhecimentos, especialmente para aqueles que estão nas extremidades da escala de proficiência, pode haver uma redução no comprimento do teste sem prejuízos na precisão da aferição.

A seleção dos itens, que conta com diferentes estratégias (Cf. BARRADA et al., 2006, 2009), irá aproximar o nível de complexidade do item ao conhecimento do(a) respondente e, mais do que proporcionar desafios possíveis ao(à) respondente, esse procedimento confere maior fidedignidade às estimativas da proficiência no TAI.

A fidedignidade ou a confiabilidade de um teste se refere à precisão dos resultados no processo de aferição. Usualmente duas teorias embasam a medida educacional: o modelo teórico de análise que enfoca o instrumento na totalidade, denominada Teoria Clássica dos Testes (TCT), e os modelos que focalizam a análise do item, denominados Teoria da Resposta ao Item (TRI). Embora a TRI não exclua a análise

pela TCT, sua perspectiva de análise ganha relevo por permitir que se estimem as proficiências na mesma escala, mesmo que o(a) respondente seja submetido à diferentes subconjuntos de itens, desde que esses itens tenham recebido parâmetros atribuídos anteriormente, característica fundamental na constituição de um banco de itens (Cf. KLEIN, 2013). Com efeito, a TRI possibilita:

a) estimar a proficiência do(a) respondente após a resolução de um conjunto de itens;
b) estimar proficiências na mesma escala, embora os(as) respondentes sejam submetidos(as) a conjuntos diferentes de itens;
c) identificar o item que minimiza o erro de medida nessa estimação.

A alínea (c) será substancial para a seleção dos itens administrados e na atribuição de maior fidedignidade ao TAI, comparativamente aos testes de papel e lápis.

Alavarse e Melo (2013a, 2013b) e Klein (2013) apontaram que a magnitude do erro de medida da proficiência estimada em um teste depende de o conjunto de itens oferecer maior informação na estimação da proficiência, aspecto relacionado ao fato de os itens apresentarem dificuldade mais ou menos aproximada à proficiência do(a) respondente. Para Klein (2013, p. 45), os “itens muito fáceis ou muito difíceis para um determinado aluno(a) fornecem pouca informação para a estimativa de sua proficiência. Itens com parâmetro de dificuldade “b” próximos da proficiência do(a) aluno(a) fornecem mais informação”. Mesmo considerando que os itens apresentados nos testes de papel e lápis são, em sua maioria, de média dificuldade, afere-se com menores erros de medida a proficiência de respondentes de desempenho médio e, consequentemente, produzem-se resultados com erros de medida que comprometem o próprio processo avaliativo de respondentes de desempenho mais baixo e mais alto (KLEIN, 2013).

Os TAI, de modo geral, requerem como componentes principais (Cf. BABCOCK; WEISS, 2012): (1) um banco de itens calibrados pela TRI, (2) uma regra para seleção do(s) item(ns) inicial(is), (3) uma regra para a seleção do item seguinte, (4) um modelo para estimar a proficiência (θ ‘*theta*’) e (5) um critério para encerrar o teste (WEISS; KINGSBURY, 1984).

Uma estrutura básica, dividida em etapas, é atribuída aos TAI (Cf. BARRADA, 2012; PITON GONÇALVES, 2012). De modo resumido, temos um critério de início do teste, com a apresentação de um item de dificuldade média (suposição de uma proficiência no centro da escala). Segue-se uma etapa de estimação da proficiência do(a) respondente com base na reação ao item, podendo ser um acerto ou um erro. Após essa estimação, será avaliado se o critério de encerramento ou de parada ou ainda de finalização do teste foi atingido. Normalmente, o critério de encerramento pressupõe o alcance de uma certa precisão na estimativa da proficiência, outros critérios podem ser usados, entre eles, atingir um número máximo e predeterminado de itens. No caso de o critério de encerramento não ter sido atingido, um novo item será selecionado, segundo um critério de seleção. Essa seleção costuma ser baseada na obtenção de um item que diminuirá o erro de medida inerente à proficiência estimada. Selecionado e administrado esse novo item, o(a) respondente terá uma nova estimação da proficiência, tendo por base a reação ao item (erro ou acerto). Diante dessa nova estimação, novamente o critério de encerramento será avaliado, constituindo um ciclo retomado a cada item selecionado e respondido. Trata-se da apresentação tão somente de um resumo das etapas, pois os inúmeros estudos sobre o TAI proporcionaram análises que resultaram em variações e novas perspectivas para cada uma delas, as quais são apresentadas na seção a seguir.

Além do destaque às etapas, o banco no qual o algoritmo seleciona os itens é de fundamental importância, visto que os critérios estabelecidos de nada servem se o banco não apresentar as características que os atendam. Para Olea, Ponsoda e Prieto (1999), um TAI depende essencialmente, mas não unicamente, de dois elementos: a existência de banco abrangente de itens, parametrizados, preferencialmente, por modelos da TRI, e um dispositivo eletrônico para seleção e apresentação do item.

Apesar de toda a potencialidade do TAI, deve-se considerar que sua construção é tema com muitos desafios a serem superados, especialmente quando exige um cuidadoso equilíbrio das contribuições da tecnologia, psicometria, *design* de testes e as ciências da aprendizagem, conforme apontam Luecht (2013) e Yan, von Davier e Lewis (2014).

Adicionalmente, apesar das vantagens dos TAI, não significa que sua utilização seja considerada a mais indicada. Wainer (2000a, 2000b) pondera restrições para seu uso,

indicando situações em que apresenta maior utilidade. Moreira Junior (2011, p. 80) considera que o TAI é mais útil nas situações em que:

> a) A natureza do construto é de tal forma que a administração informatizada ajuda na sua avaliação (por exemplo, na utilização de sons ou animação);
> b) O teste é administrado frequentemente (várias vezes ao ano);
> c) As pessoas que fazem o teste têm interesse em obter o nível de habilidade com uma alta precisão (por exemplo, identificar o nível de estresse a fim de fazer um tratamento adequado).

Esse autor também reconhece que o desenvolvimento do TAI é complexo e envolve uma grande equipe, composta por profissionais de várias áreas e recursos computacionais com *hardware* e *software* apropriados. O autor ainda assinala que no Brasil existem poucos TAI desenvolvidos.

Renom e Doval (1999), revisando outros trabalhos, corroboram as restrições apresentadas, pois mostram a existência de inconvenientes dos TAI, sobretudo pelo desequilíbrio entre o potencial teórico e o baixo impacto nas avaliações aplicadas. Destacam que esses inconvenientes podem ser quanto: ao banco de itens, aos custos, às questões técnicas e à demora da relação esforço-resultado. Quanto aos custos, os autores ponderam que uma análise no curto prazo pode levar à conclusão de que implementar um TAI é mais custoso que o teste em papel e lápis, mas, se o longo prazo for considerado, os custos de aplicações de avaliações na versão papel e lápis podem ser superiores ao de um TAI, aspecto que foi considerado na criação do CAT-AVASB.

Diante das vantagens e ponderações sobre o TAI e as preocupações em construir um dispositivo que pudesse servir de apoio para a prática avaliativa dos(as) professores(as) quanto à proficiência leitora de alunos(as) dos anos iniciais, realizou-se um levantamento bibliográfico sobre o TAI que será apresentado na subseção seguinte.

3.3 A revisão da literatura sobre o TAI

O levantamento bibliográfico sobre TAI tinha como objetivo uma aproximação do conhecimento produzido a respeito da temática. Os trabalhos revisados foram

divididos em dois tipos. O primeiro constituiu-se de obras e artigos considerados seminais por trazerem os principais aspectos sobre o TAI. O segundo agrupou teses e dissertações sobre testes adaptativos informatizados, realizadas tanto no Brasil como em outros países, disponibilizadas nas plataformas de busca utilizadas. Essa seção tem o objetivo de apresentar esse levantamento, considerando-o essencial para distinguir a produção do dispositivo no presente estudo. Como em toda revisão, é importante destacar o caráter parcial mediante o ritmo acelerado das investigações nesse campo.

3.3.1 Histórico abreviado

Os testes adaptativos são tão antigos quanto os testes psicológicos e remontam aos trabalhos sobre inteligência no início do século passado[22] (Cf. OLEA; PONSODA, 2003; PITON-GONÇALVES; ALUÍSIO, 2015). Esses testes foram construídos para adequarem-se às diferentes faixas etárias e níveis educativos, ainda que a aplicação não contasse com base computadorizada. Mesmo contando com o recurso de testes em papel e lápis, percebeu-se a possibilidade de individualizar um teste, classificando os itens de acordo com o nível de dificuldade. Essa individualização exigia que os itens fossem agrupados em subtestes segundo níveis de dificuldade. Um primeiro subteste com nível médio de dificuldade era submetido ao(à) respondente e, se todos os itens desse subteste fossem respondidos corretamente, outro subteste com itens de maior nível de dificuldade era apresentado. Caso os itens do primeiro subteste apresentado fossem respondidos de forma errada, um subteste com itens de menor nível de dificuldade era apresentado ao(à) respondente.

Olea e Ponsoda (2003) também salientam que, embora a ideia original de teste adaptativo fundamentado na TRI tenha sido proposta por Lord apenas em 1970, seu desenvolvimento iniciou já no final da década de 1960, com os estudos sobre a possibilidade de ordenar em uma mesma escala examinandos(as) que não respondiam a testes idênticos.

Bejar (2011) explica que as décadas de 1970 e 1980 do séc. XX são momentos importantes para a criação de estratégias de seleção de itens, empregadas ainda para

[22] Binet y Simon (1905, apud OLEA; PONSODA, 2003).

os testes em papel e lápis, tendo em vista que os computadores não tinham a capacidade de processamento necessário aos cálculos exigidos pela TRI. De maneira geral, eram estratégias que previam a subdivisão do teste em etapas de menor longitude, cada uma agrupava itens de um nível de dificuldade homogêneo. O(a) respondente, após passar por uma etapa de diagnóstico, era encaminhado para outra, conforme seu desempenho. Maior detalhamento das etapas pode ser encontrado em Olea e Ponsoda (2003) Renom et al. (1999) e Renom e Doval (1999). Lord (1977) também descreve teoricamente um teste adaptativo, realizado na versão papel e lápis em que o(a) próprio(a) respondente seleciona o subgrupo de itens a que irá responder.

Nesse período, em paralelo ao desenvolvimento na área da psicometria, dado pela TRI, começaram a ser aplicados os TAI em diferentes contextos. Um dos exemplos mais conhecidos de teste adaptativo é o *Test of English as a Foreign Language* (TOEFL), o teste de proficiência no inglês enquanto segunda língua, difundido em diferentes países, incluindo o Brasil. O *Graduate Record Exam* (GRE) também é um teste adaptativo realizado tanto na versão impressa como computadorizada, empregado para admissão em cursos de graduação e pós-graduação no mundo todo, embora tenha surgido nos Estados Unidos. O *Armed Services Vocational Aptitude Battery* (ASVAB), teste para seleção e classificação de candidatos ao serviço militar nos Estados Unidos, introduzido em 1968 e aplicado para mais de 40 milhões de examinandos(as), foi o primeiro a ter a versão TAI de larga escala e de alto impacto, denominado CAT[23]-ASVAB, aplicado experimentalmente de 1982 a 1984. Muitos outros TAI são aplicados em diferentes lugares do mundo, como a prova holandesa MATHCAT, que serve para avaliar conhecimentos matemáticos em pessoas adultas. Na Espanha, os primeiros TAI chegaram um pouco mais tarde, conforme elenca Barrada (2012): o TRASI, que mede a capacidade de raciocínio sequencial e indutivo, o eCAT, que mede o nível de compreensão de inglês escrito, o CAT-Health, que avalia a qualidade de vida relacionada com a saúde, e o TAI espanhol, que avalia o conhecimento da língua vasca. Testes adaptativos informatizados já foram desenvolvidos também na China, na Coreia, na Argentina e no Brasil, embora de forma muito acanhada. Chang (2012) citou outros dois exemplos de CAT em larga escala, como o *Graduate Management Admission Test* (GMAT) e o *National Council*

[23] Sigla para *Computerized Adaptive Testing* ou Testes Adaptativos Computadorizados.

of State Boards of Nursing (NCLEX). Bejar (2011) considera que os TAI verdadeiramente operativos passaram a ser administrados nos anos 1990. Luecht e Sireci (2011-12) reforçam que nas últimas quatro décadas, houve um crescimento incremental dos TBC, incluídos neles os CAT, como uma alternativa viável ao teste de papel e lápis. Esses autores também revelaram que a pesquisa inicial foi direcionada para as questões teóricas, como melhorar a eficiência de medição, atingindo níveis adequados da confiabilidade, usando o menor número de itens possível. No entanto, logo ficou evidente que também precisavam ser abordadas questões práticas, como balanceamento do conteúdo no teste, implementação de novos tipos de item e controle da exposição dos itens aos examinandos(as). Nos últimos anos, a pesquisa sobre os TAI concentraram-se no desenvolvimento dos níveis de eficiência na medição, satisfazendo simultaneamente outros objetivos importantes, como minimizar a exposição de itens e manter a validade de conteúdo. Os autores mostraram que os TAI passaram a englobar uma grande variedade de tipos de avaliação, propósitos de testes, modelos de itens e testes de desempenho, de admissão em faculdades e pós-graduações, de certificação profissional, de licenciamento, de inteligência, de linguagem, testes psicológicos, para diagnósticos médicos, de educação de adultos e de uso militar. Mudanças na tecnologia disponível também são apontadas desde os primeiros terminais conectados a um *mainframe* ou minicomputador até as estações de trabalho em rede, computadores pessoais (PC), *laptops, netbooks* e dispositivos portáteis, como *smartphones* e *tablets*.

A revisão de pesquisas e obras, apresentada nas duas subseções seguintes, permite observar que o início do século XXI mostrou uma maior abrangência nos estudos sobre o TAI, seja por meio de simulações, seja no sentido experimental.

3.3.2 O TAI nas teses e dissertações

Além das obras que tratam do assunto, foi feita uma busca por artigos, dissertações e teses indexados em acervos como o da plataforma da Fundação da Comissão de

Aperfeiçoamento de Pessoal de Nível Superior (Capes)[24], em sítios específicos de pesquisa de trabalhos científicos no Brasil, como a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD)[25], Biblioteca Digital Vérsila[26], o Portal Domínio Público[27], o Repositório Científico de Acesso aberto de Portugal (RCAAP)[28], o Sistema Integrado de Bibliotecas da Universidade de São Paulo (Sibi)[29] e o sistema de busca do Google Acadêmico.

No primeiro momento, realizou-se a busca pelas expressões “teste adaptativo informatizado”, “testes adaptativos informatizados”, “teste adaptativo”, “testes adaptativos”, “teste adaptativo computadorizado” e “testes computadorizados”, que tiveram como resultado trabalhos brasileiros, portugueses, espanhóis, mexicanos e colombianos. Após a retirada de trabalhos que não abordavam TAI, ora por abordar somente a aprendizagem adaptativa, ora por abordar testes informatizados ou computadorizados, porém não adaptativos, foi possível identificar 17 dissertações e 6 teses, identificadas em maio de 2016, além de uma tese e uma dissertação que foram acrescentadas em 2018, todas dispostas nos Quadros 2 e 3. Também foram localizados e revisados 143 trabalhos entre artigos científicos e obras sobre o TAI.

---

[24] Órgão do Ministério da Educação brasileiro responsável pelo reconhecimento e pela avaliação de cursos de pós-graduação *stricto sensu* (mestrado profissional, mestrado acadêmico e doutorado) em âmbito nacional.

[25] Sítio que reúne, em um só portal de busca, as teses e dissertações defendidas em todo o País e por brasileiros no exterior. Foi concebido e é mantido pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) no âmbito do Programa da Biblioteca Digital Brasileira (BDB), com apoio da Financiadora de Estudos e Pesquisas (FINEP), tendo o seu lançamento oficial no final do ano de 2002.

[26] Sítio que reúne e oferece gratuitamente milhões de itens digitais de produção científica oriundos dos melhores centros de pesquisa do mundo e concentra acervos abertos acadêmicos. A Biblioteca está sediada no Hemisfério Sul e foi inaugurada em novembro de 2015.

[27] O portal, lançado em novembro de 2004, coloca à disposição de todos os usuários da rede mundial de computadores – internet – uma biblioteca virtual que deverá se constituir em referência para professores, alunos, pesquisadores e a população em geral.

[28] O portal RCAAP reúne documentos de carácter científico e acadêmico, nomeadamente artigos de revistas científicas, comunicações a conferências, teses e dissertações, distribuídos por inúmeros repositórios portugueses.

[29] Portal de Busca Integrada que integra os recursos informacionais do Sistema Integrado de Bibliotecas da Universidade de São Paulo (SIBiUSP). Nessa interface única, encontram-se resultados dos recursos impressos e digitais disponíveis no Sistema Integrado de Bibliotecas.

Quadro 2 – Teses sobre testes adaptativos informatizados, por autor, ano, título, área e país

Ano	Autor	Título	Área	País
2005	MOLINA, Mª Teresa López-mezquita	*La evaluación de la competência léxica: tests de vocabulário, su fiabilidad y validez*	Linguística	Espanha
2008	KENG, Leslie	*A comparison of the performance of Testlet-Based Computer Adaptive Tests and Multistage Tests*	Filosofia	EUA
2009	MÁXIMO, Luis Fernando	*A efetividade de feedbacks informatizados sobre a auto regulação da aprendizagem em cursos a distância: um estudo de caso na área da computação*	Informática na educação	Brasil
2011	MOREIRA JUNIOR, Fernando de Jesus	*Sistemática para implantação de testes adaptativos informatizados baseados na teoria da resposta ao item*	Engenharia de produção	Brasil
2012	VIEIRA JUNIOR, Niltom	*Planejamento de um ambiente virtual de aprendizagem baseado em interfaces dinâmicas e uma aplicação ao estudo de potência elétrica*	Engenharia elétrica	Brasil
2012	PITON GONÇALVES, Jean	*Desafios e perspectivas da implementação computacional de testes adaptativos multidimensionais para avaliações educacionais*	Ciências de Computação e Matemática Computacional	Brasil
2017*	OLIVEIRA, Cassandra Melo	*Construção e busca de evidências de validade de um banco de itens de personalidade para testagem adaptativa desenvolvido a partir dos princípios do desenho universal*	Psicologia	Brasil

Fonte: Dados reunidos pela autora, em maio de 2016, em diferentes sítios de busca.
*Acrescentada em 2018.

Quadro 3 – Dissertações sobre Testes adaptativos informatizados, por ano, título e área

Ano	Autor	Título	Área	País
2001	RIBEIRO, Rui Manuel Bártolo	*Os tempos de latência nas respostas aos itens de testes informatizados: Contributos para a compreensão do processamento cognitivo*	Comportamento organizacional	Portugal
2002	OLIVEIRA, Leandro Henrique Mendonça de	*Testes Adaptativos Informatizados: uma aplicação em exames de proficiência em inglês para programas de pós-graduação*	Ciências matemáticas e de computação	Brasil
2004	PITON GONÇALVES, Jean	*A integração de testes adaptativos informatizados e ambientes de tarefas para o aprendizado do inglês instrumental*	Ciências da computação e Matemática computacional	Brasil
2009	BECHER, Ednei Luis	*Características do pensamento algébrico de estudantes do 1º ano do Ensino Médio*	Ensino de ciências e matemática	Brasil
2009	COSTA, Denise Reis	*Métodos estatísticos em testes adaptativos informatizados*	Métodos estatísticos	Brasil
2012	HOMA, Agostinho Iaqchan Ryokiti	*E-learning com análise combinatória*	Ensino de ciências e matemática	Brasil
2012	ABREU, Renata Cardoso Pires de	*Ensaio da ferramenta DIA – diagnóstico e informação do aluno*	Ciências computacionais	Brasil
2012	SASSI, Gilberto Pereira	*Teoria e a prática de um teste adaptativo informatizado*	Ciências da computação e Matemática computacional	Brasil
2013	RICARTE, Thales Akira Matsumoto	*Teste adaptativo computadorizado nas avaliações educacionais e psicológicas*	Ciências matemáticas e de computação	Brasil
2013	GALVÃO, Ailton Fonseca	*Um Modelo Inteligente para Seleção de Itens em Testes Adaptativos Computadorizados*	Ciência da computação	Brasil
2013	MOITA, Pedro Miguel Silva	*Avaliação adaptativa em dispositivos móveis das habilidades cognitivas preditoras do desenvolvimento de leitura em crianças*	Tecnologias e metodologias em *e-learning*	Portugal
2013	CASTRO, Natália Fontes Caputo de	*Tradução e adaptação transcultural do domínio fadiga do Sistema de medida de informação do relatório de resultados do paciente – PROMIS® – para a língua portuguesa*	Ciência da saúde	Brasil
2014	ARAUJO, Joacy Victor Maia	*Teoria da resposta ao item no processo de decisão*	Estatística	Brasil
2014	ALMEIDA, Caroline Medeiros Martins de	*Prática educativa usando o sistema SIENA para o ensino de ecologia no 6° ano do ensino fundamental*	Ensino de ciências e matemática	Brasil
2015	SILVA, Vanessa Rufino	*Avaliação da proficiência em inglês acadêmico através de um teste adaptativo informatizado*	Estatística	Brasil

continua

continuação do **Quadro 3.**

Ano	Autor	Título	Área	País
2015	MENEGHETTI, Douglas Rizzo	*Metodologia de seleção de itens em testes adaptativos informatizados baseada em agrupamento por similaridade*	Engenharia elétrica	Brasil
2015	MAIA JÚNIOR, Antonio Geraldo Pinto	*Uso do tempo de resposta para melhorar a convergência do algoritmo de testes adaptativos informatizados*	Estatística	Brasil
2017*	SANTOS, Jucelio Soares dos	*Mensuração de habilidades cognitivas preditoras do desenvolvimento de leitura em crianças através de jogos educacionais para dispositivos móveis*	Ciência da computação	Brasil

Fonte: Dados reunidos pela autora, em maio de 2016, em diferentes sítios de busca.
*Acrescentada em 2018.

As dissertações e teses localizadas mostraram investigações em torno de testes adaptativos computadorizados ou informatizados em diferentes direções, perspectivas e contextos.

As pesquisas de Castro (2013), Oliveira (2002) e Ribeiro (2001) focalizaram a análise das características psicométricas dos itens ou psicológicas do teste na implementação de um TAI. Essa perspectiva será denominada de validação psicométrica, pela ênfase identificada.

Ribeiro (2001) se deteve nos aspectos de validação e da análise do tempo de latência em testes aplicados na versão impressa e versão adaptativa informatizada em situações de seleção de indivíduos para admissão. Ele objetivou analisar o comportamento dos sujeitos na realização dos testes informatizados, a partir da dimensão velocidade mental, passível de medição por meio dos tempos de latência.

Oliveira (2002) propôs aumentar e calibrar um banco de itens do domínio de inglês instrumental com o propósito de obter um número ideal de itens para uso em um teste adaptativo. Ele usou o programa XCALIBRE para obter os parâmetros dos itens e desenvolveu um Sistema de Gerenciamento do Banco de Itens (SisBI) um sistema adaptativo de avaliação diagnóstica da proficiência em inglês do programa de mestrado do ICMC-USP, batizado de Testes Adaptativos para o Exame de Proficiência em Inglês (TAEPI). As funcionalidades oferecidas pelo TAEPI variam de acordo com os três tipos de usuário: administrador(a), professores(as) ou alunos(as).

O estudo de Castro (2013) traduziu e adaptou culturalmente os bancos de itens do PROMIS, um teste adaptativo informatizado norte-americano que faz a medição de

aspectos do estado de saúde obtidos diretamente do paciente, ou seja, sem a necessidade de interpretação do médico. No estudo, ela focalizou a adaptação cultural de itens do domínio de ansiedade e depressão desses bancos, mostrando a importância de assegurar que a versão traduzida esteja devidamente adaptada ao contexto linguístico e cultural da população-alvo. Embora a autora aponte alguns elementos que caracterizam um TAI para que o leitor se familiarize com a natureza do PROMIS, não aprofundou aspectos da implementação desse tipo de teste, preocupando-se com a validação do questionário traduzido.

Em direção diferente estão os estudos de Abreu (2012), Almeida (2014), Araujo (2014), Becher (2009), Homa (2012), Keng (2008), Máximo (2009), Molina (2005), Oliveira (2017), Santos (2017), Silva (2015) e Vieira Junior (2012), que estabeleceram estudos de aplicação do TAI ao diagnóstico de conhecimentos matemáticos ou ecológicos e aptidões. Eles serão identificados como aplicações do TAI, pelo enfoque presente.

Molina (2005) investigou a construção de um TAI com ênfase no vocabulário de uma segunda língua, o inglês, considerando-o como essencial para a comunicação. Para tanto, apresentou uma revisão do ensino, da aprendizagem e da avaliação do vocabulário para indicar um novo conjunto de conteúdos para o teste de vocabulário, focalizando a apresentação de TAI para avaliar o conhecimento de alunos(as) de diferentes níveis educativos na Espanha, a começar pela etapa da Educação Secundária Obrigatória (ESO), passando pelo bacharelado e finalizando com o nível correspondente ao primeiro ciclo da licenciatura universitária. Apresenta também uma versão de teste adaptativo de múltiplos estágios (MST) do 4º da ESO. O detalhado processo de construção dos testes e posterior análise de sua fiabilidade e validade permitiram concluir que os testes constituem ferramenta válida e confiável na medição padronizada da competência léxica dos(as) alunos(as) e que estes manifestam uma deficiente competência, requerendo adoção de intervenção.

Keng (2008) também estuda o desempenho de um teste na versão MST, que faz uso de *testlets* – pacotes de itens administrados juntos –, geralmente baseados em um estímulo comum. O uso de *testlets* viola a independência local, uma suposição fundamental da TRI. Comparou um TAI adaptável apenas ao nível do *testlet*, um adaptável ao nível do item e outro um MST. As condições de teste incluíram o comprimento do teste, o tamanho do conjunto de itens e a distribuição da habilidade

do(a) examinando(a). Os dados dos(as) examinandos(as) foram gerados usando parâmetros de itens calibrados pela TRI com base em dados de uma avaliação de larga escala da proficiência em leitura. Os três modelos de teste foram avaliados com base na eficácia da medição e propriedades de controle de exposição de itens. O estudo constatou que todos os três modelos adaptativos produziram uma semelhante e boa precisão de medição. O TAI no nível de item produziu melhor precisão de medição, seguido pelo *design* do MST.

Máximo (2009) fez um estudo de caso para analisar a interação de alunos(as) de um curso superior, a distância, de Tecnologia em Análise e Desenvolvimento de Sistemas com uma ferramenta para a construção de algoritmos. Essa ferramenta proporcionava *feedbacks* no módulo de avaliação adaptado ao conhecimento do(a) estudante. Ele observou que os *feedbacks* são indicativos de um processo de autorregulação da aprendizagem.

Becher (2009) estudou características do pensamento algébrico em seus(suas) próprios(as) alunos(as) do 1º ano do ensino médio. Sua pesquisa, de cunho qualitativo, implementou um *software* chamado SCOMAX, desenvolvido em projeto de pesquisa realizado conjuntamente entre Brasil e Espanha, para realização do teste adaptativo informatizado do conhecimento de álgebra desses(as) estudantes. A partir das características comuns dos erros cometidos pelos(as) estudantes, concluiu que eles(as) não desenvolveram o pensamento algébrico necessário à modelação de problemas.

Abreu (2012) também conjugou, em seu estudo, aspectos relativos à avaliação de conhecimentos matemáticos e à implementação de um TAI, denominado ferramenta DIA. A autora articula Matrizes de Referência do Sistema de Avaliação da Educação Básica (Saeb) a outros documentos, como os Parâmetros Curriculares Nacionais (PCN) para a área de Matemática e estabelece um conjunto de objetivos educacionais no qual baseou a simulação de um banco de itens e a simulação de respondentes ao teste. Ao ter como alvo um diagnóstico, o algoritmo de seleção de itens não se baseou unicamente no critério canônico de precisão da informação obtida para estimativa da proficiência, levando em consideração para a seleção do próximo item um subconjunto do banco de itens, que exclui itens cujos objetivos educacionais foram respondidos incorretamente pelo(a) respondente, ou seja, o algoritmo exclui os itens cujos objetivos são identificados como não desenvolvidos.

Homa (2012) investigou o desenvolvimento de um *e-learning*, na plataforma ILIAS, focalizando uma sequência didática que envolvia: Princípio fundamental da contagem, Permutação simples, Arranjo simples e Combinação. O experimento foi realizado com sete alunos(as) do curso de Licenciatura em Matemática, da Universidade Luterana do Brasil, no Rio Grande do Sul. Dentre os dezessete objetos de aprendizagem desenvolvidos para a sequência, um deles era o Teste Adaptativo Computacional (*iQuiz/iQuizcreate*) que permitia a autoavaliação. Os resultados apontaram que o *e-learning* é uma proposta concreta de trabalho que valoriza a autoavaliação do(a) aluno(a), com um tema matemático relevante à formação do pensamento formal para a Análise combinatória.

Vieira Junior (2012) propôs uma nova metodologia para a construção de um AVA (Ambiente Virtual de Aprendizagem), baseada nas ciências cognitivas e defendeu que esta mudança de paradigma pode favorecer os procedimentos atuais para a tecnologia educacional. Identificou modelos mentais e respectivos níveis para um dos tópicos do ensino da Engenharia elétrica, que por sua vez apresentava o maior número de *softwares* educacionais encontrados na literatura: a potência elétrica. Sobre esse tópico, propôs um algoritmo de monitoria, em tempo real, do comportamento e das características de usuários, enfocando três aspectos: pedagógico, cognitivo e de desempenho. A monitoria permitiu que as interfaces fossem dinamicamente adaptadas e individualizadas, possibilitando que características de interesse implícitas sejam gradativamente exercitadas.

Almeida (2014) também desenvolveu uma sequência didática no Sistema Integrado de Ensino e Aprendizagem (Siena) com conteúdo de Ecologia e aplicou aos(às) alunos(as) do 6° ano do ensino fundamental de uma escola municipal de Sapucaia do Sul. O SIENA é um sistema para apoio ao desenvolvimento do processo de ensino e aprendizagem de qualquer conteúdo, desenvolvido pelo convênio de pesquisa entre o Grupo de Estudos Curriculares de Educação Matemática (GECEM), da Universidade Luterana do Brasil (ULBRA) e o Grupo de Tecnologias Educativas, da Universidade de La Laguna (ULL), em Tenerife, Espanha. Para o desenvolvimento da sequência didática no SIENA, foram necessárias as seguintes ações: o desenvolvimento de um mapa conceitual com o conteúdo, a criação de uma grade dos conceitos trabalhados e, para cada conceito, uma sequência didática e um teste

adaptativo com 30 questões. Foram feitos pré e pós-testes para verificar se a aprendizagem foi efetiva, verificando as aquisições dos conceitos.

Araujo (2014) desenvolveu uma ferramenta para auxiliar os aspirantes ao serviço público a selecionarem o concurso que melhor se alinha às suas aptidões, aumentando as chances de aprovação. A ferramenta consistiu de uma testagem adaptativa computadorizada com banco de questões de concursos públicos desde 2009, totalizando 300.000 questões de aproximadamente 3.000 concursos. Ele realizou simulados e cruzou os resultados obtidos com os resultados dos concursos.

Silva (2015) descreveu as etapas, a estrutura do exame, os métodos empregados e os resultados das aplicações do teste adaptativo informatizado para proficiência em inglês (TAI-PI), que foi desenvolvido para avaliar alunos(as) do programa de pós-graduação do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP) a partir do segundo semestre de 2013. O exame adotou o modelo de resposta gradual unidimensional de Samejima, o critério de Kullback-Leibler para seleção de itens, o método de estimação pela esperança *a posteriori* para os traços latentes e a abordagem *shadow test* para imposição de restrições de conteúdo e de tamanho da prova na composição do teste de cada indivíduo. Foram apresentados os estudos de classificação em aprovados e reprovados.

Oliveira (2017) buscou construir um banco e calibrar os itens para avaliação da personalidade sob a perspectiva da Testagem Universal de pessoas com e sem deficiência, ou seja, que visam à utilização de instrumentos de forma acessível procurando abarcar a diversidade humana. A pesquisa abrangeu a conceituação do teste, a escolha do número e tipo de itens, as análises e correções, a aplicação em campo, as análises qualitativas e quantitativas e a revisão final. O teste foi implementado pela plataforma Concerto, com escolha de BIB para a organização do banco. Foram identificados cinco fatores por meio da análise fatorial e foi utilizado o modelo de resposta gradual ou de Samejima, separadamente, em cada um dos cinco fatores resultando em: 64 itens de Abertura, 58 itens de Extroversão, 82 itens de Neuroticismo, 76 de Realização e 37 de Socialização. Houve exclusão de itens após a análise do Funcionamento Diferencial dos Itens.

Santos (2017) implementou um TAI na forma de um jogo, para avaliar habilidades de aliteração, segmentação, memória visual e rima, como tarefas preditoras da leitura, no rastreio da dislexia em escolas. O jogo foi uma adaptação de um teste psicométrico

construído e validado no Brasil, denominado de Teste de Habilidades Preditoras da Leitura (THPL). Após as etapas de desenvolvimento do jogo e dos itens, calibração dos itens, elaboração do algoritmo e validação do jogo, aplicou-se o teste em escolas. O estudo ofereceu um recurso válido e rápido para levantar indícios de dislexia, auxiliando o encaminhamento de alunos(as) para especialistas.

Outra perspectiva foi observada nos estudos de Costa (2009), Galvão (2013), Maia Júnior (2015), Meneghetti (2015), Moreira Junior (2011), Piton Gonçalves (2004, 2012), Ricarte (2013) e Sassi (2012), nos quais a ênfase do estudo recai diretamente sobre aspectos teóricos e técnicos da implementação de testes adaptativos informatizados, ainda que alguns deles tratem de testes de proficiência em dada área de conhecimento. Por isso serão denominados de estudos do TAI. Piton Gonçalves (2004) verificou a integração de um TAI, que avalia o inglês instrumental para alunos(as) do mestrado em Ciências da computação e Matemática computacional do ICMC-USP, com um Ambiente Computacional de Aprendizagem (ACA), congregando aspectos pedagógicos, relativos à linguística e avaliação; computacionais, relativos à implementação do dispositivo; e estatísticos, relativos à elaboração de testes. Essa integração previa auxiliar a aprendizagem do inglês instrumental a partir dos diagnósticos fornecidos pelo TAI.

Costa (2009) comparou a implementação de três formas diferentes de estimação da proficiência em simulações do TAI também para uma prova de proficiência em Inglês instrumental I e também se preocupou em construir uma escala de proficiência com o envolvimento de professores(as) da Universidade de Brasília (UNB).

Moreira Junior (2011), além de um levantamento bibliográfico abrangente sobre TAI, desenvolveu e aplicou um método para sistematizar etapas de elaboração de um TAI, as quais podem ser utilizadas como guia de usuário. A validação do método foi realizada com a avaliação teórica para a obtenção da carteira de habilitação de motorista no Departamento de Trânsito do Estado de Santa Catarina (DETRAN-SC). Os resultados mostraram as deficiências e as vantagens potenciais da implantação de um TAI no DETRAN-SC. Considerou inédita a constituição de todas as etapas de um TAI, embora destaque que a etapa de implantação do teste foi considerada apenas na sistemática e não se efetivou por falta de recursos financeiros, físicos (computadores, provedores), humanos (programadores, *designers*) e de tempo.

Sassi (2012) mostrou como instalar e usar uma implementação de TAI desenvolvida para VBA-Excel, identificada como TAI2U, usando a interface com o R. Ele apresentou um estudo sobre os diferentes critérios de seleção de itens, testando-os em diferentes simulações; sobre modos de balancear o uso dos itens no banco; e sobre como instalar o TAI2U, que implementa três diferentes algoritmos de seleção.

Ricarte (2013), além de apresentar um detalhamento histórico sobre o surgimento dos Testes Adaptativos Computadorizados (TAC), termo usado por ele para o TAI, desenvolveu um programa para a construção de TAC, inserido no projeto Same-CAT, baseado no modelo de Samejima. Esse modelo foi usado para analisar as curvas de resposta geradas para todas as categorias de resposta, como ocorre com os modelos para itens politômicos, e não somente quanto à categoria considerada correta, como ocorre com os modelos para itens dicotômicos. Ele usou a abordagem do *shadow test* (LINDEN; HAMBLETON, 1997) na definição do algoritmo de seleção de itens. Quanto aos dados, foram usadas simulações e aplicação experimental, com alunos(as) da pós-graduação, abordando questões de inglês para o Exame de Proficiência em Inglês (EPI) do ICMC-USP, e com pacientes do Hospital das Clínicas, da Faculdade de Medicina da USP, abordando o Inventário de Depressão de Beck (BDI), nas quais foram consideradas aplicações para banco de itens grande e pequeno, respectivamente. Mostrou a necessidade de um banco abrangente para obter boas estimativas da proficiência.

Galvão (2013) propõe um modelo de seleção de itens, baseado em metas definidas para o erro padrão da proficiência resultante da aplicação do teste. Ele simulou modelos para controlar o erro local com metas de precisão a serem cumpridas a cada seleção de itens. A estratégia para a seleção de itens é baseada na previsão da variância *a posteriori*, a ser obtida caso um item seja selecionado. Salienta que os algoritmos genéricos têm como objetivo se aproximar da solução ótima para um problema, porém, o modelo de Verschoor (apud GALVÃO, 2013) propôs que o resultado atinja um valor suficiente de informação total no teste, e não o maior valor possível.

Piton Gonçalves (2012) identificou lacunas nos estudos com Teste Adaptativo baseado na Teoria da Resposta ao Item Multidimensional (MCAT) quanto: aos estudos operacionais em contextos educacionais; ao uso do critério de seleção de itens por Kullback-Leibler entre Posteriores Subsequentes (KP), que evita selecionar

um item difícil para um examinado com baixa habilidade; e às propostas de critérios iniciais e de parada. Também implementou e validou pela ISO-9126, base para avaliar testes computadorizados, uma abordagem de seleção de itens identificada por KP, um critério inicial consistente com a realidade de avaliações educacionais, discutindo a arquitetura para viabilizar uma aplicação via WEB do MCAT. Também discutiu aspectos teóricos e metodológicos da nova abordagem proposta. O KP combinado com outro critério de parada proporcionou testes mais curtos e com maior acurácia do que aqueles com a metodologia bayesiana usual, bem como tempo computacional de processamento condizente com a abordagem multidimensional.

Meneghetti (2015) propôs o agrupamento de itens, utilizando medidas de similaridade, com a finalidade de balancear duas variáveis significativas na seleção de itens do banco em um TAI: a precisão da aferição da proficiência dos(as) examinandos(as), que depende diretamente da escolha dos melhores itens possíveis a cada instante do teste; e a utilização homogênea de todos os itens disponíveis no banco de itens, sem que haja repetição. Ao utilizar o agrupamento de itens por medidas de similaridade, foram selecionados itens considerados ótimos a cada instante de seleção de itens no teste. Ele também fez a análise comparativa da aplicação de diferentes algoritmos de agrupamento em diferentes bases de itens calibradas pela TRI, por meio de simulações.

Maia Junior (2015) estruturou um modelo estatístico que considerou o tempo de resposta do item no modelo TRI, calculando-se a nova função de verossimilhança e recalculando as medidas de informações de Fisher, Kullback-Leibler e a Máxima Informação Esperada para essa nova abordagem. Ele usou simulações para comparar o modelo criado com os existentes.

Das dissertações identificadas, encontrou-se em Moita (2013) a maior aproximação com o presente estudo. Inserido no Projeto Ler, que tinha por objetivo utilizar um teste adaptativo informatizado para avaliar habilidades preditoras de leitura em crianças da faixa de 4 a 7 anos de idade, investigaram-se as vantagens que esse tipo de instrumento pode trazer, comparando as versões informatizada e papel e lápis com aplicação realizada em 300 crianças, no início do ano escolar de 2013, e aplicando questionários para três professores, 15 universitários e três professores-juízes, quanto à usabilidade da plataforma. O pesquisador concluiu seus estudos em Lisboa, mas a pesquisa empírica foi realizada em João Pessoa (Paraíba) no Brasil, em uma parceria

com a Universidade Federal da Paraíba. Embora elementos coincidentes, como a produção de um TAI envolvendo professores(as) e voltado para temática da alfabetização, o presente estudo se diferencia por enfocar um teste já existente na versão impressa e validado nacionalmente, a PB, que prevê descritores de alfabetização e letramento inicial, abrangendo aspectos cognitivos não explorados pela pesquisa de Moita (2013). Também serão analisados alunos(as) de outra faixa etária, momento em que o diagnóstico de leitura ocorre com maior frequência e para fins formativos. Este estudo também se distingue por promover a formação em avaliação de professores(as) e gestores(as) com o proporcionar participação e acompanhamento da construção do TAI, facilitando a construção de significado pedagógico para uso da plataforma e por agregar uma regra adicional ao critério de parada, com base na classificação da proficiência estimada e que possibilitará a precisão articulada ao aspecto pedagógico da escala, conforme será exposto.

3.3.3 Os artigos sobre o TAI

Além de algumas obras que tratavam do assunto, foi feita uma busca por artigos em revistas especializadas e no sistema de busca do Google Acadêmico e ERIC. Esse levantamento retornou um número de artigos muito amplo[30] e ao longo do processo abarcou-se autores recorrentemente referenciados nos artigos e volumes que tivemos acesso sem, necessariamente, optar por um recorte temporal, usualmente utilizado nas investigações. Mais ao final da elaboração do relatório, houve uma preocupação em buscar artigos que tratassem do algoritmo do TAI, especialmente, que focassem o critério de parada, para avaliar a originalidade daquele que foi usado no TAI construído. Desse modo, o levantamento nos levou a artigos como o de Weiss e Kingsbury (1984), Spray e Reckase (1996), Eggen e Straetmans (2000), Parshall et al. (2002), Mahalingam e Magis (2015), lidos na etapa final da produção deste relatório, e cujo critério de parada, ou finalização, para a classificação do(a) respondente, em certas condições, assemelham-se ao critério de parada construído para o TAI desta investigação com diferenciais que serão apresentados.

---

[30] Uma busca realizada em novembro de 2018 no Sistema ERIC pela expressão “*adaptative testing*”; retornou 365 artigos.

Foi abarcada boa parte dos artigos e obras produzidas nos polos de pesquisa sobre TAI, tendo sido identificados como principais polos de produção: os Estados Unidos, a Espanha, a Austrália, o Canadá e a China. Também foram realizados estudos no México, Argentina e Arábia Saudita, acessíveis por meio das ferramentas de busca. A leitura dos artigos, capítulos e volumes levou a um agrupamento quanto à temática considerada predominante. Isso significa que, embora um artigo ou volume tenha sido alocado em um agrupamento, ele pode abarcar aspectos pertencentes a(os) outro(s). Os estudos abordados na revisão foram identificados, conforme os itens a seguir.

3.3.3.1 Estudos iniciais sobre as características dos testes adaptativos

Nessa categoria, reúnem-se trabalhos que mostraram, seja por meio de simulação ou experimentalmente, as primeiras vantagens dos TAI em relação aos testes convencionais. São os trabalhos que surgiram ainda no século XX e construíram as primeiras evidências sobre as possibilidades desse tipo de teste.

Lord (1977), construiu duas formas paralelas de um teste personalizado de amplo alcance para avaliar habilidade verbal. O teste era apropriado para ser aplicado desde a quinta série até a pós-graduação. Administrações simuladas indicaram que o teste personalizado com 25 itens é tão bom quanto um teste convencional de 50 itens. Na maioria dos níveis de habilidade, o teste personalizado é muito melhor.

Urry (1977) mostrou que testes sob medida podem ser economicamente aplicados a problemas práticos na medição educacional. Discutiu as condições necessárias para o desenvolvimento de um sistema eficiente de testes personalizados e ilustrou aplicações de testes sob medida para habilidades únicas e múltiplas. Esclareceu que testes personalizados requerem o uso de um número muito menor de itens em comparação aos testes de papel e lápis, e que o custo tecnológico de testes personalizados não é proibitivo. Analisou as implicações dos testes sob medida para o treinamento de futuros teóricos e profissionais de medição como uma necessidade.

Kreitzberg, Stocking e Swanson (1978) mostraram que o teste adaptativo computadorizado surgiu como uma alternativa útil e vantajosa aos testes convencionais administrados por meio do papel e lápis e deveria ser considerado uma técnica de avaliação viável. Descreveram a lógica subjacente ao teste adaptativo computadorizado, discutindo os desenvolvimentos psicométricos e técnicos que o

tornavam prático, os autores também revisam alguns trabalhos anteriores que moldaram o estado atual da arte.

Kingsbury e Houser (1988) compararam as pontuações em testes convencionais e CAT[31] e concluíram que não há alterações significativas. Stocking (1997) explorou, no contexto de testes adaptativos e usando simulações, três modelos que preveem o controle da revisão de respostas pelos(as) respondentes. Dois dos modelos preservaram a imparcialidade.

Dorans (2014) discutiu a equalização na TRI e os procedimentos necessários para equalização na construção de bancos de itens em CAT.

3.3.3.2 Estudos que apresentam visões abrangentes dos testes adaptativos

Nessa categoria, foram alocados os artigos e as obras que procuraram abarcar características, estruturas, etapas, conceitos de avaliação envolvidos no TAI. São os trabalhos mais abrangentes e muitas vezes resultantes de pesquisas de um grupo de pesquisadores.

Mills (2002) reuniu os trabalhos apresentados em um colóquio patrocinado pelo *Educational Testing Service* (ETS), no qual aproximadamente 200 especialistas em medidas, pertencentes a oito países e 29 estados, se reuniram para avaliar o estado atual e futuro dos TBC. A obra se organiza em três grandes segmentos: Modelos de Teste, Administração de Testes e Análise da Pontuação do Teste. O TBC foi considerado um veículo importante para a entrega de testes e pode tornar-se a forma dominante de entrega de testes no futuro próximo. Concluíram que os TBC, incluindo os CAT, oferecem desafios para os profissionais de medição, que são necessários avanços em Psicometria, e garantem resultados confiáveis em relação às formas tradicionais de avaliação.

Martin (2003) forneceu uma breve visão geral histórica sobre o desenvolvimento dos TAI e descreveu seus componentes, lógica e implementação concreta. Em seguida, enfocou o potencial do uso de TAI como ferramenta de avaliação somativa e formativa em um contexto educacional. Para a avaliação somativa, foram apresentados

---

[31] Do termo *Computerized Adaptive Tests* (CAT)

programas de TAI em larga escala em um contexto de certificação em licenciatura, bem como problemas específicos associados a esta abordagem. Para a avaliação formativa, investigou-se se o TAI pode ser usado para orientar intervenções pedagógicas e remediativas em um contexto de aprendizagem. Foi também discutido o benefício potencial de novos formatos de itens, baseados em multimídia, e a conexão de plataformas de computador em redes.

Ponsoda Gil et al. (2004) apresentaram a revisão da pesquisa em testes adaptativos computadorizados desde 2000, data de última revisão realizada por parte de Hontangas, Ponsoda, Olea e Abad (2000, apud PONSODA GIL et al., 2004). A revisão concentrou-se nos seguintes aspectos: estimativa de habilidade, seleção de itens, elaboração e manutenção do banco de itens e a detecção de padrões de resposta inadequados.

López-Cuadrado, Pérez e Armendariz (2005) revisaram diferentes aspectos a serem considerados na decisão por recorrer aos testes computadorizados como um mecanismo de avaliação. Apresentaram a Teoria da Resposta ao Item (TRI) e as fases para a construção de um teste de avaliação, desde a concepção até a administração convencional ou adaptativa, com um banco calibrado de itens. Também trataram da avaliação em sistemas de *e-learning*.

Davey e Pitoniak (2006) propuseram uma discussão que foi necessariamente precedida por uma descrição dos métodos e procedimentos psicométricos, desenvolvidos para administrar e pontuar os testes adaptativos. Eles argumentaram que o CAT é uma boa e viável escolha apenas sob condições muito particulares. A intenção foi fornecer aos profissionais as informações necessárias para julgar, de maneira realista, as alternativas.

Sierra-Matamoros et al. (2007) descreveram a investigação teórica sobre a estrutura e o funcionamento dos TAI e suas vantagens e limitações. Também discorreram sobre os Testes Auto-adaptativos Informatizados (TAAI).

Thompson (2010) e Thompson e Weiss (2011) mostraram que o CAT é uma tecnologia de avaliação que pode beneficiar muitos programas de testagens. Os programas que exigem uma avaliação eficiente e precisa podem ser mais bem atendidos por uma abordagem CAT do que por testes tradicionais de comprimento fixo, sejam entregues via papel e lápis ou por computador. Os benefícios são:

segurança do teste, pontuações precisas e economia de tempo. Apresentou uma estrutura geral para o desenvolvimento de qualquer avaliação CAT, que deve conter: banco de itens calibrados, ponto de partida para a estimação, algoritmo de seleção de item, algoritmo de pontuação e critério de finalização.

Piton-Gonçalves, Monzón e Aluísio (2009) apresentaram dois métodos alternativos de avaliação informatizada em larga escala aplicados para estudantes de um programa de mestrado: a Medida de Probabilidade Admissível e o Teste Adaptativo Informatizado, baseado no algoritmo CBAT-2 e na TRI. Apontaram que os métodos informatizados podem apresentar vantagens em relação aos métodos convencionais.

Van der Linden e Glass (2010a) apresentaram um histórico dos testes adaptativos e suas vantagens. Argumentaram que a disseminação de avaliações em larga escala, computadores e os avanços da TRI são essenciais no desenvolvimento dos CAT. Discorreram também sobre: a seleção de itens e estimação da habilidade no CAT; restrições no CAT com *shadow test*; CAT multidimensional; programas de aplicação de testes em larga escala (MATHCAT e GRE); desenvolvimento e manutenção de bancos de itens, incluindo métodos de controle de exposição; itens inovadores; ajustes de modelos e calibração de itens; funcionamento diferencial de itens no CAT; e CAT baseados em *testlet*.

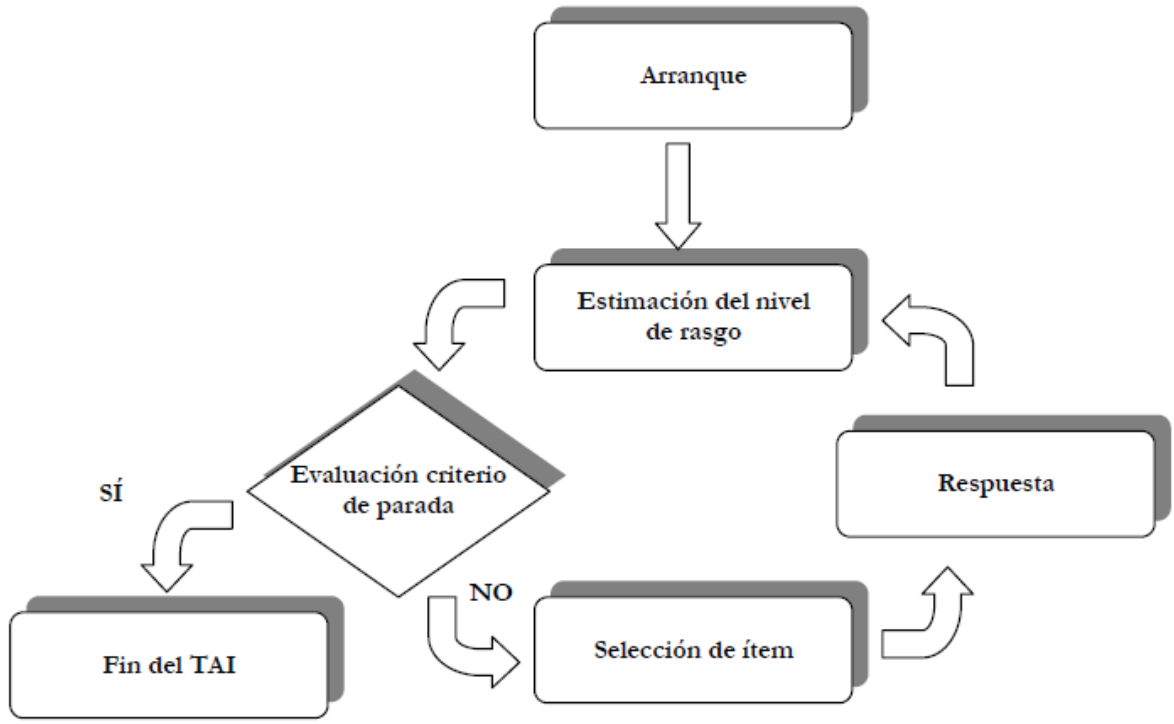
Em Van der Linden e Glass (2010b) os assuntos são retomados e a presença de novos colaboradores ampliam a abordagem dos elementos do CAT a partir da apresentação de testes japoneses; estimação de parâmetros de modelos clonados de itens (*item-cloning*); e detecção de respostas não padronizadas. O diferencial está em uma parte dedicada aos testes de múltiplos estágios, com uma revisão das pesquisas existentes e as árvores de categorias usadas no teste de classificação.

Shin, Chien e Way (2012) estudaram o balanceamento de conteúdo nos testes adaptativos computadorizados, em exames de larga escala no Canadá. Um estudo de simulação foi conduzido para comparar os métodos de desvios ponderados (WDM), de restrição associada a uma restrição CCAT e de penalidade ponderada (WPM). Três conjuntos de itens, que medem habilidades de matemática (MATH), de leitura (RD) e de um teste de licenciamento (LT), foram usados no estudo. Para o teste MATH, o WPM apresentou um desempenho ligeiramente melhor no balanceamento de conteúdo e precisão de medição. No entanto, causou a desvantagem de uma exposição de itens um pouco maior e taxas nunca utilizadas. Para o teste RD, o WPM

apresentou melhor desempenho que o WDM. Para o teste LT, os métodos WPM e CCAT tiveram 100% de taxa-alvo, enquanto o WDM apresentou algumas violações de limite inferior. A precisão da medição e os resultados do controle de exposição do item foram semelhantes nos três métodos. A opção CRWOB reduziu eficientemente o comprimento do teste de 73 para 68, para alcançar a regra de encerramento (CSEM = 0,25). O método CCAT com a opção sem bloqueio de itens não administrados (CRWOB) foi a combinação mais eficiente para o teste LT, porque exigia o menor número de itens para alcançar a mesma precisão de medida, enquanto a taxa máxima de exposição do item era controlada em torno da taxa-alvo (0,33).

Barrada (2012) buscou oferecer uma visão atualizada do TAI, apresentando a estrutura básica e as diferentes etapas que o compõem. O estudo focou na seleção de itens, parte considerada fundamental para adaptabilidade do teste, além de expor os objetivos que um TAI deve satisfazer: a) precisão; b) segurança do banco de itens; c) controle de conteúdo; e d) manutenção da prova. O artigo forneceu uma estrutura fundamental para o entendimento dos TAI, apresentado na Figura 3.

Figura 3 – Fluxograma dos TAI



Fonte: Barrada (2012).

O autor ainda identificou os objetivos básicos a cumprir em um TAI, acrescentando um aos definidos por Davey e Parshall (1995, apud BARRADA, 2012): a) permitir a

estimação precisa do nível de proficiência dos avaliados; b) limitar a probabilidade de uma filtragem de itens; c) garantir o ajuste das especificações de conteúdo da prova; d) facilitar a manutenção do banco de itens. No artigo foram apresentados vários métodos para atingir esses objetivos.

Plajner (2016), em Praga, apresentou os métodos bayesiano e de redes neurais para o CAT, que são novos na área de testes educacionais. A teoria geral associada a cada tipo é brevemente explicada, e a utilização desses modelos para CAT é analisada. Pesquisas futuras são descritas, mostrando muitos caminhos interessantes não apenas com o CAT, mas também para outras áreas da inteligência artificial.

García Jiménez, Gil Flores e Rodríguez Gómez (1998) apresentaram uma revisão do conhecimento sobre TAI e estudaram a relação entre ambientes de aprendizagem e aplicação de TAI, analisando as possibilidades de melhora dos processos de ensino e aprendizagem.

Wainer (2000a) apresentou uma introdução histórica ao volume que traz a descrição de como construir, manter e usar um sistema de testes adaptativos computadorizados. Inicia pelos primórdios dos testes mentais, relatando que alguns testes de proficiência rudimentares aconteceram na China há 2200 a.C. Em 1115 a.C., no começo da dinastia Chan, procedimentos de testes formais foram instituídos para candidatos a cargos. Os chineses descobriram que uma amostra relativamente pequena do desempenho de um indivíduo, medido sob condições cuidadosamente controladas, poderia dar uma imagem precisa de sua capacidade. Os testes mentais passaram a cumprir diferentes funções em culturas distintas. Após, surgiram os testes das forças armadas e para admissão nas universidades, nos Estados Unidos, e os testes de inteligência de Binet. Evidenciaram o surgimento da análise fatorial e das primeiras normas formais para a construção de testes. Apontaram o desenvolvimento psicométrico que proporcionou o surgimento da TRI e dos CAT, apresentando a estrutura e os componentes necessários ao seu desenvolvimento. Usaram uma aplicação CAT hipotética para ilustrar o conceito e descrever em geral as várias alternativas e variações sobre esse tema durante o volume.

Chang (2014) conduziu uma pesquisa de 18 anos de progresso na Psicometria do TAI e estabeleceu uma revisão histórica do TAI. Além disso, abordou várias questões que emergiram da implementação em larga escala e mostrou como os trabalhos teóricos podem ser úteis para resolver os problemas. Finalmente, propôs que a tecnologia TAI

pode ser muito útil para apoiar a instrução individualizada em escala de massa e mostrou que os testes baseados em papel e lápis podem ser adaptados para apoiar o ensino em sala de aula.

Magis e Mahalingam (2015) esclareceram que o recente aumento dos dispositivos eletrônicos portáteis, como *laptops, tablets* e *smartphones*, fez com que os testes adaptativos computadorizados ganhassem popularidade em face da tradicional administração de testes, administração linear e em papel, em que todos os participantes do teste recebem exatamente o mesmo conjunto de itens, possivelmente em uma ordem diferente. A combinação de testes adaptativos com avaliação baseada em computador produziu testes adaptativos computadorizados. Em vez de administrar o mesmo conjunto de itens a todos(as) os(as) examinandos(as), o teste adaptativo envolve um processo interativo, pelo qual o próximo item é selecionado para melhor estimar o nível de habilidade do(a) examinando(a) e com a maior precisão possível. Entre outros, o teste adaptativo produz testes mais curtos com menos itens, mas com informações iguais quanto à estimativa das habilidades. Além disso, as avaliações individualizadas permitem testes personalizados que, por sua vez, reduzem o risco de fraude. Esboçaram os principais passos de um processo CAT, desde o gerenciamento de um banco de itens até os processos de parada, incluindo o critério para classificação (p. 246) e também enfatizaram os *softwares* existentes (comerciais ou *open-source*). Ilustraram os códigos do *software* R para o banco de itens de um CAT.

Mahalingam e Magis (2015) apresentaram uma breve visão geral da plataforma de aplicação *on-line* e gratuita, o Concerto, fornecendo um tutorial passo a passo do desenvolvimento e da administração de um teste adaptativo.

Zhang e Chang (2016) proporcionaram uma breve visão geral das teorias e pesquisas na moderna medição e teoria dos testes, também conhecida como testes inteligentes, e ainda discutiu como esses testes se relacionaram com os conceitos e as práticas da aprendizagem inteligente. Uma introdução à aprendizagem inteligente e alguns dos desafios que ela enfrenta foram fornecidos, seguidos de uma pesquisa introdutória para alguns tópicos selecionados em psicometria, como teoria de resposta a itens, testes adaptativos computadorizados, avaliações em grande escala, diagnóstico cognitivo e equiparação. Algumas implicações a respeito dos modelos, das teorias e das técnicas de testes inteligentes na educação inteligente foram propostos,

juntamente com as descrições de alguns dos projetos em andamento e direções futuras de pesquisa.

Van der Linden (2016) analisou o desenvolvimento da metodologia Montagem de Teste Ideal (OTA), que tem como base a combinação da TRI, do banco de itens e de testes computadorizados. A TRI nos permitiria projetar e montar testes individuais adaptados ao nível atual de realização de estudantes; o banco de itens, com sua prática de redação contínua de itens, testes de campo e calibração, no permitiria criar a lista de itens necessários para adequar os testes e garantir pontuações comparáveis entre os diferentes testes; e os computadores seriam úteis para entregar testes aos(às) examinandos(as) e controlar todos os outros processos.

Lu e Cong (2016) apontaram que, nos últimos anos, testes adaptativos computadorizados estão se tornando o foco do campo da avaliação educacional. O processo de desenvolvimento de testes adaptativos informatizados foi revisto e discutido e mais recentemente foram apontados os problemas atuais, como tempo de resposta, modelo de respostas graduadas, bancos estratificados, entre outros.

Han (2018) explicou os três componentes de um algoritmo de seleção de item no CAT: balanceamento de conteúdo do teste; o critério de seleção de item; e o controle de exposição de item. Igualmente explorou as várias metodologias subjacentes a cada componente, mostrando outros fatores importantes a serem considerados ao determinar um *design* CAT adequado: os requisitos de recursos do computador; o tamanho dos *pools* de itens; e o tamanho do teste. Assumiu que a lógica do TAI está sendo adotada agora no campo da aprendizagem adaptativa, que integra o aspecto de aprendizagem e o aspecto (formativo) de avaliação da educação em uma experiência de aprendizagem contínua e individualizada. Entendeu que os algoritmos e as tecnologias descritos na revisão podem ajudar os(as) educadores(as) de saúde, médicos e os desenvolvedores de testes de alto impacto a adotarem o CAT de forma mais ativa e eficiente.

3.3.3.3 Estudos com foco nos métodos de estimação da proficiência ou de seleção de itens

Nessa categoria reúnem-se as pesquisas que calcaram na verificação dos métodos usados para estimar a proficiência dos(as) respondentes e nos critérios de seleção dos itens do TAI, seja por meio de simulação ou experimentalmente.

Bock e Mislevy (1982) mostraram que a aferição da proficiência pelo método de estimação *a posteriori* (EAP), que se baseia na avaliação numérica da média e da variância da distribuição *a posteriori*, tem propriedades extraordinariamente boas para testes adaptativos computadorizados. Nos estudos de simulação, relataram a quase equivalência do desvio padrão posterior e do erro padrão da medida.

Kingsbury e Zara (1989) apontaram os procedimentos desenvolvidos para selecionar os melhores itens para um teste adaptativo computadorizado. Analisaram várias abordagens clássicas e abordagens alternativas para a seleção de itens e discutiram seu mérito relativo.

Dodd (1990) usou conjuntos de dados reais e simulados para investigar os efeitos de duas variáveis operacionais características de testes adaptativos computadorizados. As duas variáveis estudadas foram o procedimento de seleção de itens e o método *stepsize,* utilizado até que estimativas de proficiências pudessem ser calculadas por máxima verossimilhança. Os resultados sugerem que: (a) os conjuntos de itens que consistem em apenas 25 itens podem ser adequados para CAT; (b) o método *stepsize* variável, de estimação preliminar da proficiência, produziu menos casos de não convergência que o uso do procedimento *stepsize* fixo; e (c) o procedimento de seleção de itens usado em conjunto com o critério de parada do erro padrão superou a técnica de seleção de itens usada em conjunto com uma regra de parada de informação máxima, em termos de frequências de casos não convergentes, número de itens administrados e correlações das estimativas. As implicações desses achados para a implementação do CAT com itens para avaliação em larga escala e de classificação são discutidas.

Kingsbury e Houser (1993) indicaram que a aplicação da TRI oferece avanços aos procedimentos do TAI, aprimorando as capacidades de medição e, ao mesmo tempo, alinhando os testes aos processos cognitivos dos candidatos. Esse desenvolvimento

pode melhorar a aceitação e a compreensão da mensuração pelos indivíduos que são os usuários finais dos sistemas.

Veerkamp e Berger (1997) propuseram alguns critérios de seleção de itens alternativos para testes adaptativos. Estes critérios levam em conta a incerteza da estimativa da proficiência. Sugeriram um critério de informação geral ponderada, no qual o usual critério de máxima informação e os critérios alternativos propostos são casos especiais. Um pequeno estudo de simulação foi realizado para comparar os diferentes critérios. Os resultados mostraram que o critério de informação ponderada é uma boa alternativa ao de máxima de informação.

Steinberg, Thissen e Wainer (2014) e Thissen (2014) discutiram os rápidos avanços nas tecnologias e Psicologia cognitiva e analisaram vários aspectos de testes computadorizados e adaptativos que levantam novas questões na consideração da confiabilidade da medida. Uma diferença que favorece o CAT está no fato de que a maioria dos procedimentos estatísticos utilizados no estudo de confiabilidade pressupõe que o erro de medição nos resultados dos testes seja o mesmo para todas as aplicações e isso não ocorre para as pontuações obtidas com testes de comprimento fixo. A confiabilidade está intrinsicamente ligada ao erro de medida, que vigora sob pressupostos diferentes na TCT e na TRI. Para os sistemas CAT, é possível que os erros de medição sejam aproximadamente iguais, não obstante os pressupostos da TRI.

Chang e Ying (1996) mostraram que a maioria das seleções de itens em testes adaptativos computadorizados é baseada na máxima informação de Fisher. Em cada estágio, um item é selecionado para maximizar a informação de Fisher no nível de proficiência provisoriamente estimado (θ). Entretanto, a informação de Fisher pode ser muito menos eficiente do que se presume se as habilidades estimadas não estiverem próximas do verdadeiro θ. Isso pode ocorrer especialmente em estágios iniciais de um teste adaptativo, quando o comprimento do teste (número de itens) é muito curto para fornecer uma estimativa precisa do verdadeiro θ. Argumentou-se que os procedimentos de seleção baseados na informação global devem ser usados, pelo menos nos estágios iniciais de um teste, quando as estimativas de θ provavelmente não são próximas do verdadeiro θ. Para este propósito, um procedimento de seleção de item baseado na informação global é proposto. Os resultados de estudos de simulação, que compararam a seleção de itens pela abordagem da informação

máxima e pela nova abordagem de informação global, indicam que o novo método leva a melhorias em termos de viés e redução do erro quadrático médio sob muitas circunstâncias.

Barrada et al. (2006) estudaram alguns dos problemas associados ao critério de seleção de itens baseado na máxima informação de Fisher, que tem um impacto negativo na precisão das estimativas e na segurança do banco de itens. Essa regra de seleção, implementada no e-CAT, um CAT para avaliação do nível de conhecimento do inglês escrito, é comparada com duas outras por meio de simulações: seleção do item com máxima informação de Fisher por intervalo, de Veerkamp e Berger (apud BARRADA et al., 2006), e uma nova regra, denominada informação máxima de Fisher por intervalo com média geométrica. Esse novo critério de seleção de itens fornece menor erro de medida e menores taxas de sobreposição de itens, indicando melhorias na qualidade das estimativas e na manutenção da segurança do banco de itens nos quais se baseia o e-CAT.

Chen, Lai e Mao (2007) compararam quatro mecanismos de estimação de habilidade em relação ao estado convergente e divergente e comportamentos dinâmicos sob diferentes tipos de resposta no TAI. O resultado indicou que os métodos bayesianos resultaram em convergência ou divergência lentas, mesmo esgotando o conjunto de itens para alguns padrões de resposta. Ao contrário, os métodos MLE e WLE produziram *status* convergentes para os padrões de resposta idênticos. Para superar a questão da contradição entre precisão da estimativa da proficiência e eficiência de teste, sugere-se que um sistema TAI acionado por um único mecanismo de estimativa de proficiência seja transformado nos múltiplos mecanismos de estimativa da proficiência (por exemplo, EAP + WLE). Sempre que o sistema detecta um padrão de resposta específico, o mecanismo alterna automaticamente para outro mecanismo apropriado e continua a estimar a proficiência do(o) examinando(a).

Kim Kang e Weiss (2007) usaram a simulação Monte Carlo para comparar quatro procedimentos de estimação da proficiência em CAT. Focaram em ambientes de aprendizagem flexíveis em que as diferenças individuais entre os(as) alunos(as) são importantes. A medida levaria em consideração o quanto o nível de desempenho de um determinado estudante, medida da mudança individual (AMC), foi alterado como resultado de experiências instrucionais em um currículo definido.

Clares López (2008) fez uma proposta de organização dos itens em blocos para a forma de seleção de itens em um teste informatizado, cujo propósito é reduzir de uma maneira inteligente o número de itens e o tempo de resolução, entre outros elementos.

El-alfy e Abdel-aal (2008) propuseram uma nova abordagem, que usa modelagem de redes, para identificar automaticamente o subconjunto mais informativo de itens de um teste, usado para avaliar efetivamente os(as) examinandos(as) sem degradar seriamente a precisão. Os resultados indicaram que a abordagem proposta reduz significativamente o número de itens necessários, mantendo uma qualidade aceitável do teste.

Barrada et al. (2009a) compararam várias regras de seleção de item utilizadas em TAI: máxima informação de Fisher, máxima informação de Fisher considerada para o intervalo, informação de Fisher ponderada com a função de verossimilhança, informação de Kullback-Leibler, considerada em um intervalo e Kullback-Leibler ponderada com a função de verossimilhança. Esta última mostrou uma maior precisão na estimativa das proficiências nas fases iniciais do TAI. Foi proposta a informação de Fisher por intervalo, com média geométrica, que foi a única regra que superou os problemas detectados com uso da regra da máxima informação de Fisher considerada para o intervalo. Para as demais regras, reconheceu a existência de um *trade-off* entre precisão e segurança.

Chen (2009) estudou o problema de como construir CAT em modelos de classes latentes (LCM) e apresentou a aplicação da metodologia de seleção sequencial ótima na seleção de itens, construída com base em modelos de diagnóstico cognitivo. Duas novas heurísticas foram propostas e comparadas com o método de seleção de itens randomizados e com as duas heurísticas investigadas em Xu et al. (apud CHEN, 2009). Finalmente, mostrou a conexão entre as abordagens baseadas em informações de Kullback-Leibler e a baseada em entropia de Shannon (apud CHEN, 2009), bem como a conexão entre algoritmos construídos sobre o LCM e aqueles construídos sobre modelos de TRI.

Rulison e Loken (2009) discutiram o potencial de subestimação de estudantes com alta habilidade em CAT. Relataram que a maioria das pesquisas sobre estimativas imparciais de θ se concentra na identificação de aberrações nos padrões de resposta (VAN KRIMPEN-STOOP; MEIJER, 2000, apud RULISON; LOKEN, 2009) ou na adaptação de algoritmos de seleção de item, por exemplo, estratificar o *pool* de itens

e usar itens menos discriminantes no início do teste (CHANG, 2004, apud RULISON; LOKEN, 2009; CHANG; YING, 1996; PASSOS et al., 2007, apud RULISON; LOKEN, 2009). Os autores argumentaram que pode ser de utilidade para o CAT outra abordagem que envolva o modelo da TRI de quatro parâmetros, proposta por Barton e Lord (1981, apud RULISON; LOKEN, 2009).

Chen (2010) apresentou um novo método de seleção de itens, o Índice Máximo de Discriminação Global Modificado (MMGDI) para testes adaptativos computadorizados de diagnóstico cognitivo (CD-CAT). O índice modificado capturou dois aspectos de um item: (1) quantidade de contribuição que ele pode dar para uma cobertura adequada de cada atributo; e (2) quantidade de contribuição que ele pode dar para recuperar o perfil latente cognitivo. O estudo de simulação demonstrou que o método é capaz de garantir uma cobertura adequada de todos os atributos medidos pelo teste. Além disso, em comparação com o método do Índice de Discriminação Global (GDI) original, melhorou a taxa de recuperação de cada atributo e de todo o perfil cognitivo. Essa variação não ocorre quando o número de aplicações do TAI é alto. Barrada et al. (2010) compararam a eficiência de regras de seleção de itens em testes adaptativos computadorizados, propondo uma estratégia para realizar uma comparação dessas regras. A estratégia foi aplicada em um estudo de simulação de CAT de comprimento fixo para a comparação de seis regras de seleção de itens: o ponto de informação Fisher, informação de Fisher ponderada pela verossimilhança, Kullback-Leibler ponderado pela verossimilhança, máxima informação estratificada com restrição e os métodos progressivo e proporcional. Os resultados mostraram que não há regra ótima para qualquer valor de sobreposição ou erro quadrático médio (RMSE). Os melhores métodos foram o Kullback-Leibler ponderado pela verossimilhança, o método proporcional e o método de máxima informação estratificada com restrição.

Sassi e Curi (2010) estudaram vários critérios de seleção de itens (máxima informação, bayesianos, sequencial de Owen, pré-*posteriori*) para TAI. No estudo da simulação, foi usado o modelo logístico de três parâmetros da TRI, e como estimador do conhecimento/habilidade foi usada a média *a posteriori*. As simulações foram realizadas a partir de três tamanhos distintos de bancos de itens.

Weiss (2011) se preocupou em mostrar diferenças na precisão das medidas obtidas em testes computadorizados totalmente adaptativos e os convencionais. Referiu que

o processo de construção de um teste convencional de comprimento fixo frequentemente se concentra em maximizar a confiabilidade e a consistência interna, selecionando itens que são de média dificuldade e alta discriminação para o teste (um teste de “pico”). Visto do ponto de vista da TRI, as pontuações desses testes são precisas apenas para examinandos(as) cujos níveis de proficiência estão próximos do ponto de pico do teste, mas a precisão diminui substancialmente à medida que as proficiências dos(as) examinandos(as) desviam da média. Por outro lado, os escores dos testes adaptativos têm erros de medida mais homogêneos nos diferentes níveis em que a proficiência é estimada. Conceitos básicos de testes adaptativos são introduzidos e testes baseados em TRI são descritos. Vários exemplos de registros de respostas nos CAT[32] foram discutidos para ilustrar como funcionam. Alguns problemas operacionais, incluindo exposição de itens, balanceamento de conteúdo e itens inimigos, também foram brevemente discutidos. Concluiu-se que, uma vez que o CAT oferece um teste exclusivo para o(a) examinando(a), os resultados dos CAT serão mais precisos e deverão fornecer melhores dados para as pesquisas e aplicações em Ciências Sociais.

Li et al. (2011) apresentaram um teste adaptativo baseado na TRI com o objetivo de medir com precisão o potencial dos candidatos. Eles descreveram o algoritmo e o modelo da TRI utilizado na estimação da proficiência dos candidatos.

Moreira Junior (2013) estudou, via simulação, cinco algoritmos baseados na máxima informação, para estimar o grau de usabilidade de *sites* de *e-commerce* via TAI. O algoritmo com o melhor desempenho foi aplicado nos dados reais de 361 *sites* e conseguiu-se obter uma boa estimativa do grau de usabilidade com a aplicação de 13 itens.

3.3.3.4 Estudos sobre controle da taxa de exposição de itens e preocupações com a segurança do teste

Em um TAI, os itens mais informativos do banco ou *pool* de itens têm maior probabilidade de serem selecionados, pois irão garantir maior precisão na estimação da proficiência. Devido a essa característica, itens específicos do conjunto de itens

[32] Optou-se por usar a abreviação CAT tanto para referir-se ao teste no singular como no plural.

podem ser apresentados com muita frequência e se tornar superexpostos, enquanto outros itens raramente são selecionados pelo algoritmo e ficam subexpostos. Os itens superexpostos podem ser memorizados e, desse modo, comprometer a estimação da proficiência, visto que o(a) respondente acerta sem necessariamente dominar a habilidade de que trata o item, e a segurança do teste. Nessa categoria, encontram-se as pesquisas com preocupações referentes à segurança do teste bem como as que buscam minimizar a exposição excessiva de itens no banco.

Stocking e Lewis (1998) se interessaram em garantir a segurança do *pool* de itens no ambiente de testes contínuos, possibilitados pela administração computadorizada de um teste, em oposição ao ambiente de testes periódicos, afetos aos testes lineares de papel e lápis. Apresentaram um novo método de controle da taxa de exposição de itens condicionada ao nível de habilidade no ambiente de teste contínuo. As propriedades de tal controle condicional das taxas de exposição de itens foram exploradas por meio de estudos simulados e usadas em conjunto com um algoritmo de teste adaptativo.

Revuelta e Ponsoda (1998) propõem dois novos métodos para controle de exposição de itens. No método Progressivo em que, à medida que o teste avança, a influência de um componente aleatório na seleção do item é reduzida e a importância da informação do item é cada vez mais proeminente. No método de Máxima Informação Restrita, nenhum item pode ser exposto em uma proporção maior que a predeterminada nos testes. Ambos os métodos foram comparados com seis outros métodos de seleção de itens em relação à precisão do teste e às variáveis de exposição do item. Os resultados mostraram que o método de Máxima Informação Restrita foi útil para reduzir as taxas máximas de exposição e que o método Progressivo reduziu o número de itens não utilizados. Ambos se saíram bem na precisão. Um método que combina Progressivo-restrição pode ser útil para controlar a exposição do item sem uma redução séria na precisão do teste.

Revuelta, Ponsoda e Olea (1998) apresentaram três novos métodos para controlar as taxas de exposição de itens em testes adaptativos computadorizados. O primeiro adiciona um componente aleatório ao método de informação máxima. No segundo, a taxa de exposição de cada item é controlada diretamente e apenas para os itens cuja taxa de administração é menor que um determinado valor. O terceiro método é uma mistura dos dois anteriores. Os três métodos são testados por simulação, juntamente

com o método de informação máxima e o de Sympson e Hetter. Os resultados são estudados em termos da precisão do teste e controle da taxa de exposição. As vantagens de cada método são discutidas e algumas linhas de pesquisa são sugeridas.

Eggen (2001) estudou o problema da alta exposição de alguns itens do banco de itens de um TAI. Assinalou que a superexposição de itens pode ser eliminada com o acréscimo de restrições aos métodos de seleção, mas adverte que isso afeta a eficiência do TAI. Apresentou uma solução para ambos os problemas com base em simulação, que foi realizada para desenvolver testes adaptativos.

Chen e Doong (2003) queriam encontrar uma fórmula que descrevia a relação entre parâmetros de exposição de itens e parâmetros de itens em testes adaptativos computadorizados, usando programação genética (GP), uma técnica de inteligência artificial inspirada biologicamente. Os resultados mostraram que uma fórmula interessante entre os parâmetros de exposição do item e os parâmetros do item em um *pool* pode ser encontrada usando GP. Os parâmetros de exposição previstos com base na fórmula encontrada estavam próximos aos observados no procedimento de Sympson e Hetter (1985, apud CHEN; DOONG, 2003). Resultados semelhantes foram observados para o modelo multinomial de Stocking e Lewis (1998, apud CHEN; DOONG, 2003) para seleção de itens e o procedimento de Sympson e Hetter com balanceamento de conteúdo. A abordagem GP proposta forneceu uma solução para encontrar a taxa exposição em relação aos parâmetros dos itens.

Chang e van der Linden (2003) conduziram estudos sobre o procedimento de seleção de itens no CAT de Chang e Ying (1999, apud CHANG; VAN DER LINDEN, 2003), que estratificaram o banco conforme os valores dos parâmetros dos itens e exigiram que os valores dos parâmetros *b* sejam distribuídos uniformemente por todos os estratos. Assim, os valores dos parâmetros *a* e *b* devem ser incorporados nos estratos formados. Um refinamento foi proposto, baseado na estratificação de itens de Weiss (1973, apud CHANG; VAN DER LINDEN, 2003) de acordo com valores de *b*. Estudos de simulação utilizando um banco de itens retirados do exame de *Graduate Record* (GRE) indicaram que a nova abordagem melhorou o controle das taxas de exposição de itens e reduziu os erros quadráticos médios.

Hau e Chang (2001) salientaram que a segurança do teste se tornou um problema, pois os itens de alta discriminação têm maior probabilidade de ser selecionados e

superexpostos. Parece haver uma compensação entre a alta eficiência nas estimativas de capacidade e o uso equilibrado de itens. Apresentaram quatro estudos com dados simulados, concentrando-se na noção de que itens com menor poder de discriminação deveriam ser usados primeiro no TAI. O primeiro estudo demonstrou que o método de máxima informação com o controle de Sympson e Hetter (1985) resultou na seleção de itens com maior poder de discriminação no início do TAI. Os demais estudos apontaram para o uso de itens na ordem inversa, ou seja, itens com menor poder de discriminação primeiro, como descrito no método estratificado de Chang e Ying (1999, apud HAU; CHANG, 2001), tinha vantagens potenciais: um uso de itens mais balanceado e um uso relativamente estável para o *pool* de itens resultantes, com gerenciamento fácil e barato. Esse método estratificado pode ter uma eficiência melhor ou próxima a de outros métodos na estimação da proficiência, principalmente para *pools* de itens operacionais, quando os itens retirados não podem ser totalmente reabastecidos com itens semelhantes e altamente discriminativos. Argumentou-se também que a seleção criteriosa de itens, como no método estratificado, proporciona um controle mais ativo da exposição dos itens, igualando com sucesso o uso de todos os itens.

Chang e Zhang (2002) estudaram a taxa de sobreposição de itens para um grupo de examinandos(as), um problema para o CAT administrado a pequenos grupos de examinandos(as) em intervalos de tempo frequentes. Pretendeu-se evitar que os(as) respondentes submetidos ao teste antes compartilhem informações com os examinandos(as) que farão o teste mais tarde, diminuindo o risco dos itens se tornarem conhecidos. Para um *pool* de itens específico, diferentes algoritmos de seleção de itens podem gerar taxas de sobreposição distintas. Uma questão importante no projeto de um bom algoritmo de seleção de item é manter a taxa de sobreposição de um nível predefinido. Ao fazer isso, é fundamental investigar qual seria a taxa mais baixa para todos os possíveis algoritmos de seleção de itens. Os autores provam que, se cada item tem uma possibilidade igual de ser selecionado no CAT de comprimento fixo e o número de sobreposição de itens para quaisquer amostras aleatórias α de examinandos(as) segue a família distribuição hipergeométrica para $\alpha \geq 1$, os valores esperados para o número de itens sobrepostos entre a amostra aleatória α de examinados podem ser calculados com precisão. Esses

valores podem servir como referência no controle das taxas de sobreposição de itens para testes adaptativos de comprimento fixo.

Chang e Ansley (2003) compararam as propriedades de cinco métodos de controle de exposição de itens usados para estimar as habilidades dos(as) examinandos(as) em um contexto CAT. Cada um dos algoritmos de controle de exposição foi incorporado ao procedimento de seleção de item do teste adaptativo. Os méritos e as deficiências dessas estratégias foram considerados sob diferentes tamanhos de conjuntos de itens e distintas taxas máximas de exposição desejadas e foram avaliados à luz da segurança do teste, da taxa de exposição de itens e dos erros-padrão de medição condicionais. Por fim, forneceram diretrizes para a escolha de métodos apropriados de controle da exposição dos itens, de modo que a preocupação com a segurança do teste no ambiente CAT pudesse ser diminuída.

Van der Linden (2003) estudou o método de Hetter e Sympson (1985, 1997, apud VAN DER LINDEN, 2003) de controle probabilístico de exposição de itens em testes adaptativos computadorizados. Esclareceu que a definição dos seus parâmetros de controle para valores admissíveis requer um processo iterativo de simulações computacionais que foi considerado demorado, particularmente se os parâmetros tiverem que ser definidos em um conjunto realista de valores e condicionados ao parâmetro de proficiência dos(as) examinandos(as). São identificadas propriedades formais do método para explicar a razão do processo iterativo ser lento e não garantir a admissibilidade de introduzir alternativas a ele. O comportamento dessas alternativas foi estudado para um teste adaptativo de um *pool* de itens do Law School Admission Test (LSAT). Duas dessas alternativas mostraram um comportamento atraente e convergiram suavemente para a admissibilidade de todos os itens em um número relativamente pequeno de etapas de iteração.

Chen, Ankenmann e Spray (2003) apresentaram uma forma analítica de derivar matematicamente as taxas de sobreposição de teste em função da taxa de exposição do item, para testes adaptativos computadorizados de comprimento fixo. Esta relação algébrica foi usada para investigar o controle simultâneo da exposição do item tanto no item quanto nos níveis de teste. Os resultados indicaram que, em CAT de comprimento fixo, os métodos de controle de exposição de item que implementam uma especificação máxima, por exemplo Sympson e Hetter, forneceram o controle mais direto tanto no item quanto nos níveis de teste.

Van der Linden e Veldkamp (2004) implementaram o controle de exposição de itens em CAT pela imposição de restrições de inelegibilidade de itens no processo de montagem dos testes sombra (*shadow test*). O método se assemelhou ao método de controle de exposição de itens de Sympson e Hetter em que as decisões para impor as restrições são probabilísticas. No entanto, esse método não requeria estudos de simulação demorados para definir valores para os parâmetros de controle antes do uso operacional do teste. Em vez disso, foi possível definir as probabilidades de inelegibilidade do item de forma adaptável durante o teste, utilizando as taxas reais de exposição do item. Um estudo empírico, utilizando um *pool* de itens do *Law School Admission Test,* mostrou que a aplicação do método resultou em controle perfeito das taxas de exposição do item e teve impacto insignificante nas funções de viés e erro quadrático médio da estimativa da proficiência.

Chen e Lei (2005) propuseram um método de controle de exposição de item, que é a extensão do procedimento Sympson e Hetter e pode fornecer controle de exposição de item tanto no item quanto nos níveis de teste. A taxa de exposição do item e a taxa de sobreposição do teste são dois índices comumente usados para rastrear a exposição de itens em testes adaptativos computadorizados. Considerando os dois índices, a exposição do item pode ser monitorada nos níveis de item e teste. Para controlar a taxa de exposição do item e testar a taxa de sobreposição simultaneamente, o procedimento modificado tentou controlar não apenas o valor máximo, mas também a variância das taxas de exposição do item. Os resultados indicaram que a taxa de exposição do item e a taxa de sobreposição do teste puderam ser controladas simultaneamente por meio da implementação do procedimento modificado. O controle de exposição do item foi melhorado e a precisão da estimativa da proficiência diminuiu quando uma taxa de sobreposição máxima pré-especificada foi rigorosa.

Barrada, Mazuela e Olea (2006) proporcionaram aumento na segurança em TAI e estudaram o método estratificado (AS) de Chang e Ying (1999, apud BARRADA; MAZUELA; OLEA, 2006), o qual define para o início do teste apenas a administração de itens com parâmetros de discriminação *a* baixos, com os valores dos parâmetros aumentando durante o teste. Com esse método, a distribuição das taxas de exposição dos itens é menos distorcida, enquanto a eficiência da estimativa é mantida. O parâmetro de pseudo-adivinhação *c*, presente no modelo logístico de três parâmetros,

é considerado irrelevante e não é utilizado no método AS. O modelo *Maximum Information Stratified* (MIS) incorpora o parâmetro *c* na estratificação do banco e na regra de seleção de itens, melhorando a precisão em comparação com o AS, para bancos de itens com parâmetros *a* e *b* correlacionados e não correlacionados. Para ambos os tipos de bancos, os métodos de bloqueio de *b* (CHANG; QIAN; YING, apud BARRADA; MAZUELA; OLEA, 2006) melhoram a segurança do banco de itens.

Barrada, Olea e Ponsoda (2007) estudaram o método Sympson-Hetter (SH), usado como um meio de controlar a taxa máxima de exposição de itens em TAI, no qual, por meio de uma série de simulações, são definidos parâmetros de controle que marcam a probabilidade de administração de um item ao ser selecionado. Considerando que esse método apresenta dois problemas principais – requer um longo tempo para calcular os parâmetros e a taxa de exposição máxima é ligeiramente acima do limite fixo –, os autores também estudaram as alternativas que parecem resolver ambos os problemas, propostas por van der Linden (2003, apud BARRADA; OLEA; PONSODA, 2007). Foi mostrado que esses métodos restringem em demasia a exposição de alguns itens altamente discriminantes e, assim, a precisão é diminuída. Também é mostrado que, quando a taxa de exposição máxima desejada está próxima do valor mínimo possível, esses métodos oferecem uma taxa máxima de exposição empírica claramente acima da meta. Um novo método foi apresentado e se baseia na estimativa inicial da probabilidade de administração e na probabilidade de seleção dos itens com o método restrito (REVUELTA; PONSODA, apud BARRADA; OLEA; PONSODA, 2007). Esse novo método, quando usado com o SH e com os dois métodos de van der Linden, aceleraram a convergência dos parâmetros de controle sem diminuir a precisão.

Georgiadou, Triantafillou e Economides (2007) mostraram que a questão do controle de exposição recebeu maior atenção após os pesquisadores reconhecerem as vantagens dos CAT sobre os testes convencionais. Os autores apresentam uma revisão das estratégias de controle de exposição do item que têm sido apresentadas na literatura no período de 1983 a 2005, com o objetivo de prevenir a superexposição de alguns itens e aumentar a taxa de uso de itens raramente ou nunca selecionados.

Veldkamp e van der Linden (2008) demonstraram que o método Sympson-Hetter (SH) para controle de exposição em testes adaptativos pode ser implementado com a abordagem de restrição *shadow test* (STA). Os autores propõem modificações e

melhorias no método original. O método foi aplicado a um teste adaptativo com 433 restrições em vários atributos. Tanto uma abordagem única como múltipla do *shadow test* foram utilizadas para comparar listas de diferentes tamanhos para o método SH. Dessa forma, dois requisitos foram necessários para a implementação bem-sucedida do método SH: um número suficiente de itens para a seleção e o balanceamento de conteúdo durante todo o teste. O método SH com base no STA também é aplicável quando há um grande número de especificações a ser cumpridas e reduziu consideravelmente a superexposição dos itens populares no *pool* de itens. Como resultado do controle de exposição, 435 dos 753 itens do banco tornaram-se ativos durante o teste. Na perspectiva custo-benefício, o número ainda é baixo, mas maior do que os 10% ou mais dos itens em geral usados em testes adaptativos, sem o controle de exposição. Citaram outras fontes de solução para o problema.

Barrada, Olea e Abad (2008) examinaram estratégias para evitar que os(as) examinandos(as) antecipem parte do conteúdo de um teste adaptativo computadorizado, e seus níveis de proficiência estimados tenham um viés positivo acentuado. Uma das estratégias consiste em dividir um grande banco de itens em vários sub-bancos e alternar o sub-banco empregado (VELDKAMP; VAN DER LINDEN, apud BARRADA; OLEA; ABAD, 2008). Afirmaram que a estratégia permite melhorias substanciais no controle de exposição a um custo pequeno na precisão da medida, no entanto, não sabemos se esta opção oferece melhores resultados do que usar o banco mestre com maior restrição nas taxas máximas de exposição (SYMPSON; HETTER, apud BARRADA; OLEA; ABAD, 2008). Em simulações, trabalharam com vários bancos e 2100 itens, comparando RMSE e taxa de sobreposição de itens, com os mesmos bancos divididos em dois, três até sete sub-bancos. Por meio da manipulação extensiva da taxa máxima de exposição em cada banco, descobriram que a opção de alternar sub-bancos superou ligeiramente a opção de restringir a taxa máxima de exposição do banco principal, por meio do método SH.

Barrada, Olea, Ponsoda e Abad (2008) estudaram a melhoria da exposição de itens no TAI, opondo duas regras de seleção, os métodos progressivos (REVUELTA; PONSODA, apud BARRADA; OLEA; PONSODA; ABAD, 2008) e o método proporcional (SEGALL, apud BARRADA; OLEA; PONSODA; ABAD, 2008) à regra de seleção de item mais comumente usada: a máxima informação de Fisher para o nível da proficiência estimada. O método de Fisher compreende uma distribuição altamente

desequilibrada das taxas de exposição do item, uma alta taxa de sobreposição entre os exames e, para o gerenciamento do banco de itens, uma alta pressão para substituir itens com um parâmetro de alta discriminação. Uma alternativa para esses problemas envolve basear na aleatoriedade a seleção de itens no início do teste. Depois, conforme o teste progride, o peso da informação na seleção aumenta. Diferentes funções que definem o peso do componente aleatório de acordo com a posição do item a ser administrado no teste foram estudadas. As funções foram testadas em bancos de itens simulados e em um banco operacional. Concluíram que os métodos progressivo e proporcional toleraram um alto peso do componente aleatório com perda mínima ou zero de precisão, enquanto a segurança e a manutenção do banco são melhoradas.

Barrada, Abad e Veldkamp (2009) também apresentaram estudos sobre três métodos para controlar a taxa máxima de exposição de itens em TAI e indicaram que os três métodos: a) podem ser interpretados como métodos para construir a variável sub-banco de itens a partir da qual cada examinando(a) recebe os itens de seu teste; e b) possuem limitações teóricas e empíricas, comparando seu desempenho. Com os três métodos, obtiveram resultados basicamente indistinguíveis na taxa de sobreposição e RMSE (diferenças na terceira casa decimal). Aconselharam o uso do método de elegibilidade de itens, pois economiza tempo e satisfaz as metas de restringir a exposição máxima.

Em Barrada et al. (2009b, 2011), encontrou-se a preocupação com o risco de os(as) examinandos(as) receberem estimativas de proficiências inflacionadas devido ao conhecimento prévio do item, ao considerarem que testes contínuos com um mesmo banco de itens possibilita que futuros examinandos(as) obtenham informações de examinandos(as) anteriores sobre os itens que receberam. Explicaram que taxas de sobreposição mais baixas ou distribuições mais homogêneas de uso dos itens podem não levar a TAI mais seguros. Mostraram, em três estudos diferentes nos quais a divulgação do banco de itens é simulada, que há probabilidade de conhecer antecipadamente os primeiros itens administrados e que as taxas de sobreposição para os altos níveis de proficiência avaliados são melhores para a segurança do teste. Esses estudos comparam a máxima informação de Fisher, o método progressivo e os métodos de seleção alfa-estratificados e concluíram que o método de estratificação alfa não foi o método de seleção que ofereceu o nível mais alto de segurança de teste.

Chajewski e Lewis (2009) fizeram um estudo com o objetivo de simular as condições encontradas em CAT que avaliavam habilidades não cognitivas, otimizando o controle de exposição de itens em avaliações que utilizavam bancos de itens relativamente pequenos, preservando, ao mesmo tempo, a precisão da estimativa dos traços latentes. Apresentaram revisão bibliográfica sobre controle de exposição. Usaram o modelo politômico da TRI. Compararam três mecanismos de controle de exposição do item (Sympson-Hetter condicional, restrição progressiva da informação máxima e taxa total de exposição simplificada) contra duas condições de restrição do banco de itens na administração do TAI (informação máxima simples e seleção aleatória de itens). Os métodos concorrentes foram avaliados em sua capacidade de fornecer controle de exposição e viés adequado e apresentaram benefícios e deficiências, que são discutidos.

Barrada, Abad e Olea (2014) se preocuparam com a segurança em testes adaptativos computadorizados também focando nas diferentes regras de seleção de itens propostas para aliviar esse risco. Argumentaram que os métodos estratificados estão entre aqueles que receberam mais atenção. Nesses métodos, apenas itens de baixa discriminação podem ser apresentados no início do teste e a informação média dos itens aumenta conforme o teste continua. Para isso, o banco de itens deve ser dividido em vários estratos de acordo com as informações dos itens. No estudo, os autores simularam as condições das taxas de exposição com números diferentes de estratos e chegaram à conclusão de que a melhor opção é estratificar o maior número possível.

Ferrão e Prata (2014) conduziram um estudo de simulação baseado no método do critério mínimo de erro-variância, variando a taxa de exposição do item (0,1; 0,3; 0,5) e o comprimento máximo do teste (18, 27, 36). A comparação é feita examinando-se o viés absoluto, a raiz quadrada do erro médio e a correlação. Testes de hipóteses são aplicados para comparar as distribuições reais e estimadas. Os resultados sugeriram a redução considerável do viés à medida que se aumentou o número de itens administrados, observaram a ocorrência de efeito teto em testes de tamanho muito pequeno e a concordância total entre distribuições verdadeiras e empíricas para testes computadorizados de comprimento menor que os testes em papel e lápis.

Ozturk e Dogan (2015) investigaram os efeitos de diferentes métodos de controle de exposição de itens (randômico, Sympson-Hetter e *Fade-Away*) sobre a precisão da medição e a segurança do teste, considerando ainda vários métodos de seleção de

item e características de *pool* de itens. O método *Fade-Away* para controlar a exposição de itens produziu melhores resultados do que os demais.

Foster (2015) ressaltou que os TBC e CAT ao mesmo tempo que introduziram algumas preocupações com a segurança, resolveram outras. Mostrou que as vantagens obtidas com o uso da tecnologia são significativas, incluindo pontuação imediata, locais de teste mais convenientes, testes mais eficientes, melhor avaliação de habilidades e custos mais baixos. Esses avanços ocorrem em um momento em que as novas ferramentas tecnológicas também possibilitaram fraudes em testes e a *internet* ajudou a trapacear e compartilhar amplamente o conteúdo do teste. As vantagens da tecnologia para testes superam as desvantagens e agora é possível usar *smartphones* e *tablets* como parte do cenário da avaliação. O aumento do uso de testes educacional reflete na responsabilidade em buscar e manter uma boa segurança de teste. Uma visão geral das ameaças à segurança associadas a testes baseados em tecnologia e um modelo para configurar uma defesa eficaz são oferecidos.

3.3.3.5 Estudos que verificaram diferenças nas aplicações papel e lápis versus *TBC, incluindo a calibração e pontuação de itens*

Um dos pressupostos para um TAI é a existência de um banco de itens parametrizados por métodos estatísticos como a TRI. Muitas vezes, esse banco foi parametrizado por um teste em papel e lápis e, dessa forma, seria necessário verificar se haveria vieses no comportamento do(a) respondente em relação à administração por meio digital. Os artigos desse agrupamento estudaram os aspectos envolvidos na administração de testes via computador e as consequências para os parâmetros dos itens.

Renom, Doval e Sélles (1998) investigaram as possibilidades empíricas envolvidas na elaboração de um TAI com uso do procedimento de autoarranque (PA), uma rotina para a construção de banco de itens (BI) elaborada por e implementada no programa DEMOTAC2 (RENOM; MARTÍNEZ, 1994, apud RENOM; DOVAL; SÉLLES, 1998), que permite calibrar os itens do banco sem a necessidade de realizar ancoragens ou equiparações, evitando o inconveniente de um processo complicado e dispendioso, pois exige trabalhar com grandes amostras.

Sawaki (2001) detalhou que a informatização de testes de leitura L2 tem sido de interesse para os pesquisadores de avaliação linguística nos últimos 15 anos, mas poucos estudos empíricos avaliaram a equivalência do construto medido em testes de leitura L2 computadorizados e convencionais e a generalização dos resultados de leitura computadorizada para outras condições. A fim de abordar várias questões que cercam o efeito do modo de apresentação sobre o desempenho no teste de leitura L2, analisam a literatura em testes de capacidade cognitiva e medição educacional e psicológica, entre outros. A generalização dos resultados para a avaliação computadorizada de L2 foi considerada difícil por diversos motivos: a natureza das habilidades avaliadas na literatura de avaliação não envolveu necessariamente dados de linguagem; o modo de apresentação de estudos na literatura de não-avaliação envolvendo leitores de L2 é escasso; e existem limitações nos métodos de pesquisa utilizados. No entanto, a literatura levanta questões importantes a serem consideradas em estudos futuros sobre o modo de apresentação em avaliação de linguagem.

Wang e Kolen (2001) resumiram três categorias gerais de critérios para avaliar a comparabilidade entre CAT e testes de papel e lápis (PPT): (1) validade, (2) psicometria e (3) suposição estatística/administração de teste. Esses critérios também são aplicáveis para avaliar a comparabilidade entre os testes computadorizados lineares e o PPT. Para um CAT, os procedimentos de avaliação podem se tornar mais complicados devido a vários problemas relacionados aos procedimentos de administração do CAT, como estimação de parâmetro de item, seleção de item, pontuação de teste e regra de parada. A essência do critério de validade é examinar se os construtos medidos pelas versões de teste alternativas são os mesmos. Satisfazer esse critério é mais desafiador para um CAT, porque a administração de itens de maneira personalizada faz com que as diferenças de conteúdo e as diferenças entre os(as) examinandos(as) sejam mais prováveis. Diversas técnicas foram desenvolvidas para avaliar a dimensionalidade, mas a avaliação direta da dimensionalidade em CAT necessita de mais pesquisas. Devem ser realizadas investigações sobre a comparabilidade entre as versões de teste alternativas e quando um CAT é adotado junto com um PPT, devem ser examinados tanto o modo e os efeitos de paradigma, quanto as questões relacionadas à adaptabilidade.

Clariana e Wallace (2002) confirmaram vários fatores-chave na avaliação baseada em computador *versus* avaliação em papel e lápis. Com base em pesquisas anteriores,

os fatores considerados no estudo incluíram familiaridade com o conteúdo e familiaridade com o computador, competitividade e gênero. Após a instrução em sala de aula, os(as) alunos(as) de graduação de negócios (N = 105) foram aleatoriamente designados para um teste baseado em computador ou um teste idêntico em papel. A análise de variância (ANOVA) dos dados mostrou que o grupo do teste baseado em computador superou o grupo de teste baseado em papel. Gênero, competitividade e familiaridade com computadores não estavam relacionados a essa diferença de desempenho, embora a familiaridade com o conteúdo estivesse. Estudantes com melhor desempenho se beneficiaram mais da avaliação baseada em computador. Com o atual aumento na avaliação baseada em computador, instrutores e instituições devem estar cientes e analisar os possíveis efeitos do modo de administração do teste.

Bjorner (2004, 2005) apresentou diretrizes para desenvolver bancos de itens para o TAI, considerando o potencial de melhora significativa na avaliação dos resultados de saúde. Esta revisão descreveu os recursos exclusivos de bancos de itens e CAT e discutiu como desenvolver bancos de itens. Defendeu que o desenvolvimento de um banco de itens é um processo em múltiplos estágios que requer uma definição clara do construto a ser medido, bons itens, uma cuidadosa análise psicométrica dos itens e uma especificação clara do CAT final. A análise psicométrica precisa avaliar os pressupostos do modelo de TRI, como unidimensionalidade e independência local; que os itens funcionam da mesma maneira em diferentes subgrupos da população; e que há um ajuste adequado entre os dados e os modelos de resposta ao item escolhidos. Além disso, diretrizes de interpretação precisam ser estabelecidas para ajudar na aplicação clínica da avaliação.

Kim e Huynh (2007) compararam as pontuações dos(as) estudantes obtidas em testes no formato convencional e computadorizados para exames de larga escala estaduais de fim de curso (EOC), nas disciplinas de Álgebra e Biologia. Evidências indicaram que não foram encontradas diferenças significativas nos modos de administração, especialmente quanto ao construto a ser medido.

Al-Amri (2008) também comparou testes de leitura L2 baseados em papel e em computador, analisando o impacto das características dos candidatos (estudantes de Medicina da Arábia Saudita), como a familiaridade com o computador, atitudes no computador, preferência de modo de teste e estratégias de teste, sobre o

desempenho deles. Embora tenham encontrado uma diferença significativa entre médias das pontuações nas duas versões do teste, nenhuma das características dos examinados teve influência no desempenho dos(as) alunos(as) ao fazer os testes baseados em computador.

Karkee e Fatica (2010) compararam as pontuações entre modos de administração de testes (papel e lápis e *on-line*). Estudaram os efeitos do ajuste do modelo, o funcionamento diferencial do item e a média dos parâmetros de itens e pessoas. Os resultados do teste não mostraram efeitos estatisticamente discerníveis no ajuste do modelo, DIF ou desempenho do(a) aluno(a), apesar de algumas diferenças nos parâmetros do item.

Costa e Ferrão (2015) apresentaram evidências estatísticas da correlação entre a TCT e os modelos da TRI, as quais poderiam contribuir para o baixo custo na calibração de itens no contexto de sistemas de avaliação adaptativa. A TCT e o Modelo Generalizado de Crédito Parcial se aplicaram a testes que são formados por itens de múltipla escolha, respostas curtas, preenchimento de espaços ou respostas abertas, parcialmente qualificadas. Os conjuntos de dados foram extraídos dos testes aplicados à população portuguesa composta por estudantes do 4º e 6º anos. Os intervalos de confiança obtidos pelas amostras de *bootstrap* evidenciaram uma forte relação entre as estimativas da dificuldade do item e, portanto, corroboraram de forma significativa a teoria estatística dos testes mentais, mesmo com as ferramentas sofisticadas de hoje.

Chen, Liu e Ying (2015) consideraram a calibração de novos itens para a manutenção de bancos de itens de avaliações de larga escala. Fazendo uso dos estimadores de máxima verossimilhança e bayesianos para estimar as proficiências, os autores propuseram dois métodos para a calibração. Esses métodos são aplicáveis tanto a testes tradicionais, baseados em papel e lápis, para os quais a seleção de itens é pré-fixada, quanto para testes adaptativos computadorizados, para os quais a seleção de itens é sequencial e aleatória. Simulações abrangentes foram feitas para avaliar e comparar o desempenho dessas abordagens. Extensões para outros modelos de classificação diagnóstica também são discutidas.

Way et al. (2015) discutiram alguns problemas e considerações relacionados à comparabilidade das pontuações em testes administrados em computador e em papel e lápis. Também apresentaram uma rica revisão da literatura, que evoluiu ao longo

das últimas quatro décadas, ao que se refere a essa comparabilidade, considerando também a comparação entre testes em condições de ajustes, adaptações para estudantes com necessidades especiais, ou não. Examinaram vários problemas emergentes na comparabilidade, em particular aqueles associados a rápidos avanços na tecnologia, tais como uso de itens inovadores e a administração usando dispositivos digitais. Apresentaram alguns princípios e estratégias para testes de programas de larga escala.

Bejar (2011) estudou uma abordagem baseada na validade para garantir a qualidade das pontuações de itens de resposta construída em meios automatizados. A pontuação automatizada de respostas construídas já está operacional em testes de vários programas e a demanda aumenta em ritmo acelerado. Concluiu que um dos elementos-chave é o *design* de motores de pontuação, a extração de evidências e a síntese de evidências, discutidos como um novo caminho para a manutenção da qualidade das pontuações.

Williamson, Mislevy e Bejar (2006) forneceram exemplos sobre sistemas de pontuação automatizada válidas para avaliações TBC, que utilizam tarefas inovadoras. A ênfase da obra está nos objetivos de medição e como os métodos estatísticos e a tecnologia interagem para atingir esses objetivos. Também trataram do apoio aos que pesquisam domínios de avaliação orientados para o desempenho, por exemplo, prática médica, operação de sofisticados equipamentos mecânicos e elétricos e engenharia. O volume abordou a história de inovações tecnológicas aplicadas à pontuação automatizada de testes de múltipla escolha; os princípios do *Design* Centrado na Evidência (ECD), projetos que garantem a validade de construto de uma avaliação, técnicas de automação da pontuação que dependem de métodos bem conhecidos com uma história relativamente extensa. Uma visão futura dos métodos automatizados de pontuação para tarefas de resposta construídas em testes baseados em computador é oferecida.

Abad et al. (2010) estudaram a deterioração dos parâmetros estimados dos itens em uma aplicação no formato papel e lápis e em um TAI. Os resultados mostraram que há variação nos parâmetros estimados na aplicação informatizada, sendo que essa variação afetou principalmente a aferição da competência em inglês das pessoas de altas proficiências.

3.3.3.6 Estudos que analisam novos formatos de itens e de tarefas nos TAI

Com a administração de testes em meio digital, novos formatos de itens e a possibilidade de pontuação automatizada de itens de resposta construída surgem no cenário dos TBC e TAI. Os estudos com esse enfoque foram agrupados neste tópico.

Graudina e Grundspenkis (2006) descreveram uma tentativa de estender a avaliação adaptativa do conhecimento do(a) aluno(a) para tarefas baseadas nos mapas conceituais. As regras de como gerar mapas conceituais a partir de ontologias também foram propostas, bem como sugestões de como usar ontologias para avaliação do mapa conceitual criado pelo(a) aluno(a). O modelo conceitual da arquitetura do sistema foi descrito.

Sireci e Zenisky (2006, 2016) mostraram que os testes baseados em computadores permitem formatos inovadores de itens (incluindo gráficos, vídeo e áudio de alta resolução). Os autores revisam formatos inovadores de itens computadorizados, com foco nas implicações de tais tecnologias na validade da pontuação do teste. Exemplos de programas de testes que usaram itens tecnologicamente inovadores são apresentados e discutidos. Forneceram ilustrações de vários formatos de itens computadorizados e concluíram com críticas a alguns formatos de itens com relação à validade e aos critérios práticos.

Scalise e Gifford (2006) apontaram que a tecnologia oferece novas oportunidades para inovação em avaliação educacional por meio de novas tarefas de avaliação e pontuação potencialmente poderosa, relatórios e mecanismos de *feedback* em tempo real. Uma limitação em potencial para obter os benefícios da avaliação baseada em computador, tanto na avaliação instrucional quanto nos testes em larga escala, está em projetar questões e tarefas com as quais os computadores podem efetivamente interagir (ou seja, para pontuar e propor relatório final) enquanto ainda obtém evidências significativas de medição. O artigo introduziu uma taxonomia envolvendo 28 tipos diferentes de itens inovadores, que pode ser útil na avaliação baseada em computador. Organizada de acordo com o grau de restrição das opções do(a) respondente para responder ou interagir com o item ou tarefa de avaliação, a taxonomia proposta descreve um conjunto de tipos de itens icônicos, denominados itens de “restrição intermediária”. As respostas a esses tipos são consideradas totalmente restritas (ou seja, a questão convencional de múltipla escolha) e respostas totalmente construídas (ou seja, de ensaio tradicional), o que pode ser um desafio

para os computadores analisarem significativamente, mesmo com as ferramentas sofisticadas de hoje. Os 28 tipos de exemplos de itens discutidos estão ordenados em 7 categorias, envolvendo respostas com restrições sucessivamente decrescentes, de totalmente selecionadas para totalmente construídas. Cada categoria de restrição inclui quatro exemplos icônicos. O propósito foi fornecer um recurso prático para desenvolvedores de avaliação, bem como uma estrutura útil para a discussão de formatos e usos inovadores de avaliação em ambientes baseados em computador.

Shermis e Burstein (2013) analisaram os métodos e as tecnologias mais recentes usados na avaliação automatizada de ensaios[33] (AEE), incluindo pontuação automatizada de ensaios e *feedback* de diagnóstico. Os destaques incluem o que há de mais recente na avaliação do desempenho na escrita, baseados nos recentes avanços do ensino de escrita, testes de linguagem, psicologia cognitiva e linguística computacional. A obra incluiu as últimas pesquisas sobre avaliação automatizada de ensaios; descrições dos principais dispositivos de pontuação, incluindo o *e-rater*®, o *Intelligent Essay Assessor*, o *IntelliMetric* ™ *Engine*, o *c-rater*™ e o *Light*SIDE; aplicações dos usos da tecnologia, incluindo um sistema de larga escala; uma estrutura sistemática para avaliar a pesquisa e os resultados tecnológicos; e descrições de métodos AEE, que podem ser replicados para outros idiomas, além do inglês, como em um exemplo da China.

Reckase (2015) comentou que o desenho, o desenvolvimento e a pontuação de testes têm sido considerados mais arte do que ciência. Esclareceu que o uso do item de múltipla escolha torna a pontuação do teste mais objetiva, mas esse formato fala pouco sobre os processos cognitivos que os(as) examinandos(as) usam para chegar à resposta. Salienta que os laboratórios de conhecimento cognitivo e itens de respostas abertas[34] têm sido usados para obter mais *insights* sobre os processos de respostas a itens de teste, mas parece que as informações obtidas nesses laboratórios não são transferidas para o processo de desenvolvimento de itens. As seções comentadas relataram algumas tentativas de tornar o desenvolvimento de itens, a construção de testes e a pontuação de itens abertos mais científicos ou pelo menos mais reprodutíveis.

---

[33] Ensaios são as questões de resposta construída, também identificadas como questões dissertativas.

[34] Significa o mesmo que itens de resposta construída.

3.3.3.7 Estudos sobre o desenvolvimento e aperfeiçoamento de um TAI

O TAI vem sendo estudado por décadas e sua implantação já é uma realidade. Assim, o presente agrupamento contém estudos experimentais do TAI.

Moreno, Segal e Hetter (1997) descreveram as etapas concluídas e futuras do desenvolvimento do CAT-ASVAB, a versão adaptativa do teste de seleção para as forças armadas dos Estados Unidos.

Wise e Kingsbury (2000) ofereceram uma série de desafios práticos que devem ser enfrentados pelos profissionais de medição que querem implementar e manter um programa de testes adaptativos. Discutiram quatro tipos gerais de desafios: estabelecer e manter *pools* de itens, escolher procedimentos de administração de teste, proteger a segurança do teste e responder aos problemas do(a) examinando(a). Recomendaram que o sucesso de um programa de testes adaptativos dependerá em grande parte de quão bem o praticante de medição lida com esses desafios.

Olea et al. (2004) mostraram um teste adaptativo computadorizado que avalia o nível de conhecimento do inglês escrito para falantes de espanhol. Descreveram a elaboração do banco de itens, a verificação de suas propriedades psicométricas, o ajuste obtido para o modelo logístico de três parâmetros e as principais características do algoritmo adaptativo. Em um estudo de simulação, forneceram evidências sobre as propriedades dos níveis estimados de inglês (precisão e viés), analisando características psicométricas como taxa de exposição, métodos e modelos.

Ho e Yen (2005) apresentaram um experimento para examinar os efeitos sobre precisão e eficiência de uma plataforma de administração de CAT. Cinquenta estudantes do ensino médio foram selecionados para ser avaliados quanto ao vocabulário de língua inglesa tanto no computador pessoal (PC) como no assistente pessoal digital (PDA), o que lhes permitiu comparar as vantagens relevantes e as desvantagens das duas plataformas de administração. Ambos os testes usaram o mesmo banco de itens e calibrações, algoritmo de estimativa da proficiência e estratégia de seleção de itens. Os resultados indicaram que as plataformas não afetam o desempenho dos(as) examinandos(as) no CAT. As respostas do questionário sobre os ambientes de testes também mostraram que a maioria dos(as) examinandos(as) prefere fazer o teste no PDA. Concluiu-se que o uso de um PDA

para administrar CAT é tão preciso e eficaz quanto um PC e mais agradável e conveniente.

Santos e Guedes (2005) abordaram uma ferramenta computacional para realização de avaliação adaptativa em cursos à distância, utilizando como fundamentação a TRI e os testes adaptativos informatizados. Propuseram dois modelos avaliativos (MA). O MA I seleciona o item seguinte logo após um item ter sido respondido; se o(a) aluno(a) respondeu corretamente, será fornecido um novo item com maior nível de dificuldade e se o(a) aluno(a) respondeu incorretamente, o próximo item a ser administrado será um de dificuldade inferior. O MA II, após o(a) examinando(a) iniciar o teste e responder corretamente o primeiro item, seleciona o próximo item a ser administrado calculando a média entre o nível do item atual e o nível máximo admitido no teste. O MA I e II iniciam o teste com item no mesmo nível de dificuldade. Os resultados mostraram que a ferramenta desenvolvida tem a capacidade de fornecer subsídios para a instituição, para o(a) professor(a) e para o(a) próprio(a) estudante, mostrando informações que auxiliam no julgamento das estratégias individuais de estudo. Observaram que o modelo II trouxe melhores resultados para os(as) estudantes de altas habilidades.

Aguilar e Kaijiri (2007) se propuseram a desenvolver uma ferramenta TAI denominada Sistema de Avaliação Baseada em Computação (SPEBC), que gera tarefas ou atividades para serem realizadas em sala de aula de Química para o ensino médio no México, para apoiar professores(as) na realização de avaliação formativa. A ferramenta inclui uma variedade de estratégias de avaliação, tais como: conhecimento e inventário de Estudo Prévio (KPSI) (Cf. TAMIR; LUNETTA, apud AGUILAR; KAIJIRI, 2007), perguntas que começavam com o que, quando etc. e ensaios. Concluíram apontando alguns desafios a serem enfrentados no desenvolvimento da SPEBC como a geração de perguntas e respostas personalizadas com base no grau de dificuldade, seu *design* e sua implementação. Indicaram que pesquisas devem ser focadas na conclusão da implementação de o primeiro protótipo gerador de questões.

Harms e Adams (2008) apresentaram considerações sobre o *design* e a usabilidade no desenvolvimento de produtos baseados em computador, como a avaliação e o aprendizado. Também enfatizaram as ferramentas de gerenciamento de informações usadas para, além de gerenciá-las, distribuí-las. A aplicação do *Design* Centrado no Usuário (UCD) como uma ciência do *design*, ou seja, um paradigma de pesquisa direcionado a compreender, avaliar e documentar sistematicamente os requisitos de

uso, deve ser considerado no contexto de projetar e implantar ferramentas de avaliação e aprendizado baseadas na WEB.

Baylari e Montazer (2009) propuseram um sistema personalizado de agentes, baseado na Teoria da Resposta ao Item (TRI) e na rede neural artificial (RNA), que apresentam testes adaptativos baseados em TRI e recomendações personalizadas, baseadas em RNA. Esses agentes adicionam adaptatividade e interatividade ao ambiente de aprendizado e atuam como um instrutor humano que orienta os(as) alunos(as) em um ambiente de ensino amigável e personalizado na WEB.

Olea, Abad e Barrada (2010) mostraram o considerável desenvolvimento de testes adaptativos computadorizados, as características, as vantagens e desvantagens e os exemplos concretos de outros tipos de testes demandados no âmbito da Psicologia como: a) os testes baseados em modelos (disponibilizam um modelo ou teoria de como cada item é respondido, o que permite prever sua dificuldade); b) os testes ipsativos, que são eficazes no controle de alguns vieses de resposta em avaliações profissionais; c) os testes comportamentais, que medem características que normalmente são medidas com autorelatos ou por meio de tarefas que requerem respostas não verbais; e d) os testes situacionais, nos quais uma situação de conflito no trabalho é apresentada ao avaliado.

Olea et al. (2011) descreveram o *e*CAT-*Listening*, um novo teste adaptativo informatizado para a avaliação do inglês. Foram descritos o banco de itens, desenho e propriedades psicométricas do banco de itens, no qual a calibração compreendeu 1.576 participantes. Boas garantias psicométricas são evidenciadas.

Davey (2011) mostrou que o termo CAT é aplicado uniformemente em uma família diversificada de métodos de teste, que embora compartilhem objetivos e meios semelhantes, diferem de maneiras que podem afetar significativamente o desempenho dos testes quando entregues. Os autores fornecem perguntas que os responsáveis por sistemas CAT podem usar para desvendar as diferenças entre os usos. As respostas a essas perguntas podem ajudar a julgar se a experiência necessária foi de fato exercida.

Lozzia e Attorresi (2009) apresentaram os passos seguidos no desenho de um TAI aplicado com o *software* FastTEST Pro para avaliar analogias verbais em estudantes de Psicologia.

Magis e Raîche (2012) descreveram um *framework* de testes adaptativos computadorizados e apresentaram um Pacote R, chamado catR, para a simulação de padrões de resposta. Este pacote requer um banco de itens, previamente calibrado de acordo com o modelo logístico de quatro parâmetros (4PL) ou qualquer modelo logístico mais simples. O pacote propõe vários métodos para selecionar os itens do início do teste, vários métodos para a seleção do próximo item, estimadores diferentes de habilidade (máxima verossimilhança, Bayes modal, esperada *a posteriori*, verossimilhança ponderada) e três regras de parada (com base na duração do teste, na precisão das estimativa da proficiência ou classificação do(a) estudante).

Manseira e Misaghi (2013), baseados na inexistência de relatos de usos abrangentes de Testes Adaptativos Computadorizados na realidade diária de estudantes de cursos de educação a distância (EaD), conduziram o desenvolvimento de um módulo para a gestão de aprendizagem e para a realização de exercícios e avaliações formativas.

Pommerich, Segall e Moreno (2009) relataram que o CAT-ASVAB detém a distinção de ser a primeira bateria de testes adaptativos de larga escala a ser administrada em um ambiente de alto risco, são aproximadamente 20 anos de desenvolvimento e 20 anos em administração operacional. Atualmente, em torno de dois terços dos militares realizam o CAT-ASVAB, com planos de substituição total do teste em papel e lápis por CAT. Os autores também traçaram a progressão do CAT-ASVAB por meio de nove fases principais de desenvolvimento, incluindo: pesquisa e desenvolvimento do protótipo CAT-ASVAB, o desenvolvimento inicial de procedimentos psicométricos e *pools* de itens, implementação operacional inicial e em escala total, a introdução de novo *pools* de item, a introdução da administração no *Windows*, a introdução da administração na *internet* e pesquisa e desenvolvimento da próxima geração CAT-ASVAB. Apresentaram um histórico e discussões de grandes questões operacionais e de pesquisa, abordagens e práticas inovadoras e lições aprendidas para cada fase.

Fries (2014) usou o teste adaptativo computadorizado para aprimorar a estimativa da capacidade funcional de uma pessoa, com base em respostas ao banco de itens do Sistema de Informação de Medição de Resultados (PROMIS), do Instituto Nacional de Saúde dos EUA. Concluíram que as medidas foram mais precisas.

Veldkamp e Matteucci (2013) propuseram o uso de antecedentes empíricos para diminuir os custos envolvidos no CAT. Introduziram métodos para extrair *priors* empíricas baseadas nas variáveis dos candidatos e dos itens, aumentando a

eficiência do CAT. Uma estrutura geral e as várias etapas do CAT foram apresentadas, seguidas do modelo bayesiano de CAT com os procedimentos para elucidar *prior* empíricas e estimar parâmetros dos itens.

Wang et al. (2016) reconheceram que a maioria dos projetos de CAT e MST exibe forças e fraquezas em implementações recentes em grande escala, não havendo uma resposta simples para a pergunta sobre qual projeto é melhor, pois modos diferentes podem se adequar a situações práticas distintas. Propuseram uma estrutura adaptativa híbrida para combinar CAT e MST, inspirada em uma análise da história do CAT e do MST. O procedimento proposto é um *design* que transitou de um *design* sequencial de grupos de itens para um *design* totalmente sequencial de itens. Isso permitiu a robustez do MST nos estágios iniciais, mas também compartilhou as vantagens do CAT nos estágios posteriores, com o ajuste fino na estimação da proficiência, uma vez que sua vizinhança tenha sido identificada. Os resultados da simulação mostraram que os *designs* híbridos, seguindo os princípios propostos, forneceram precisão ou eficiência de estimativa comparável ou melhor do que os modelos CAT e MST padrão, especialmente para examinandos(as) nas duas extremidades do intervalo de variabilidade da proficiência.

Nunes et al. (2015) descreveram em um breve histórico, as principais etapas e componentes de um CAT, enfatizando a calibração real de um banco de itens para avaliação da personalidade por meio do XCalibre versão 4.1. Foi utilizado o CATSim para as simulações do banco. Esse banco foi adaptado para aplicação por meio do Concerto.

Spenassato et al. (2015) utilizaram o teste adaptativo computadorizado para avaliar a maturidade do Sistema de Gestão Ambiental (SGA) de empresas. O teste realizado convencionalmente foi considerado longo pelas empresas e continha questões relativas à política ambiental, ao planejamento, à implementação e operação, à verificação e ação corretiva e à análise crítica pela administração para a melhoria contínua. Uma simulação foi realizada para avaliar o efeito de diferentes regras de finalização do TAC[35] e os resultados foram comparados com a sua versão convencional. Com o TAC, poderia haver uma redução de 71% no comprimento do

[35] Abreviação para a expressão “Testes Adaptativos Computadorizados”, equivalente à tradução da expressão “*Computerized Adaptive Test*” empregada na língua inglesa.

teste sem comprometer a validade e precisão da medida, beneficiando tanto os(as) respondentes quanto os responsáveis pelo desenvolvimento e aplicação dos testes.

3.3.3.8 Estudos do TAI de múltiplos estágios, avaliação baseada em jogos, modelo de diagnóstico cognitivo e multidimensional

Os TAI podem se diferenciar quanto à forma de personalização do teste. Essa personalização pode ser realizada após a administração de um conjunto de itens ou após a administração de um único item. Também se diferenciam quanto ao modelo psicométrico utilizado na análise dos itens. Os estudos que discutiram esses assuntos estão agrupados neste tópico.

Keng et al. (2010) estudaram o teste adaptativo múltiplos estágios (MST), por considerar as deficiências nas questões de segurança do TAI. Ao contrário da maioria das pesquisas em testes adaptativos que foi baseada na TRI, os testes com uso de *testlets*, que são pacotes de itens administrados juntos, geralmente são baseados em teoria da resposta aos *testlets* (TRT), tendo em vista que o uso de *testlets* viola a independência local, uma suposição fundamental do TRI. Compararam três tipos de testes e constataram que todos produziram uma boa precisão na medição.

Huebner (2010) mostrou que a modelagem diagnóstica cognitiva se tornou um novo campo empolgante de pesquisa psicométrica. Esses modelos visam diagnosticar o *status* de domínio dos(as) examinandos(as) de um grupo de habilidades ou atributos discretamente definidos, fornecendo assim informações detalhadas sobre seus pontos fortes e fracos. A combinação do diagnóstico cognitivo com avaliações adaptativas por computador surgiu como um novo campo. O estudo forneceu uma introdução e uma visão geral dos desenvolvimentos recentes em avaliações adaptativas computacionais de diagnóstico cognitivo.

Chang (2012) apresentou uma variedade de métodos psicométricos que podem ser utilizados para montar sistemas de CAT, como as ferramentas de diagnóstico para escolas de ensino fundamental e médio a fim de classificar os níveis de domínio dos(as) alunos(as) para um determinado conjunto de habilidades cognitivas necessários para que os(as) alunos(as) tenham sucesso. No estudo, várias questões foram discutidas: como selecionar um projeto de custo econômico de *hardware* e rede para escolas?; como incorporar a função de diagnósticos cognitivos em um algoritmo

de seleção de itens?; como obter controle mais eficiente sobre restrições não psicométricas?; como equilibrar conteúdo e controle de exposição de itens? e assim por diante. Além disso, foram fornecidas informações e discussões com relação aos fundamentos psicométricos de dois projetos CAT; um é o CAT regular, que tem sido usado por mais de três décadas, e o outro é o CAT recém-surgido de diagnóstico cognitivo (CD-CAT). Finalmente, alguns resultados promissores de implementações em larga escala de CD-CAT na China foram relatados sobre o tema da aplicabilidade dos métodos propostos em ambientes K-12.

Piton-Gonçalves e Aluísio (2015) apresentaram parte dos achados envolvidos em Piton Gonçalves (2012) referentes ao TAC Multidimensional (MCAT), baseado na Teoria de Resposta ao Item Multidimensional (MIRT). A revisão da literatura aponta que o MCAT é adequado para testes computadorizados com múltiplas habilidades, administrando um número menor de itens do que os testes tradicionais.

Martín-Fernández et al. (2016) exploraram um teste adaptativo de múltiplos estágios para avaliar a inteligência fluida (FIMT) em estudantes universitários. O banco foi calibrado no modelo de resposta graduada para itens de resposta construída e foram produzidas duas estruturas de múltiplos estágios para estudo das propriedades psicométricas quanto à potencialidade da informação. Os resultados desses dois estudos apoiaram o emprego da FIMT, uma ferramenta que utiliza esse formato para avaliar de forma inovadora e precisa a inteligência fluida.

Mislevy et al. (2015) abordaram tópicos psicométricos, como modelos e validade, situando a discussão na arena da avaliação baseada em jogos (GBA). O objetivo é conectar os conceitos e métodos de avaliação e psicometria com os dos jogos.

Luecht (2013, 2015) e Luecht e Sireci (2011-12) apresentaram as principais características dos CBT e dos CAT e analisaram o cenário atual e tendências futuras. Nos textos de 2011-12 e 2015 incluíram o TAI de múltiplos estágios e os *testlet*. Mostraram que há uma tendência para plataformas de código aberto.

Kim, Moses e Yoo (2015a, 2015b) investigaram a precisão dos estimadores de proficiência da TRI em testes de múltiplos estágios (MST). Vários painéis MST de dois estágios (módulos) foram montados em cada módulo, com diferentes níveis de dificuldade e comprimento do módulo. Para cada situação, investigaram a precisão na estimação da proficiência em diferentes formas de estimação. Os estimadores

bayesianos foram ligeiramente mais eficientes que os estimadores não-bayesianos, resultando em menor erro total. Possíveis alterações de pontuação, causadas pelo uso de diferentes estimadores de proficiência, seriam não-negligentes, particularmente para os(as) examinandos(as) de baixo e alto desempenho.

Yan e von Davier (2014) e Magis, Yan e von Davier (2017) forneceram uma visão geral prática e da teoria sobre testes adaptativos computadorizados (CAT) e testes de múltiplos estágios (MST) e ilustraram metodologias e aplicações usando a linguagem de código aberto R, oferecendo vários exemplos. Mostraram que a implementação pode depender dos pacotes R catR e mstR que já foram ou estão sendo desenvolvidos e que incluem alguns dos mais recentes algoritmos de pesquisa sobre o tema.

Kimura (2017) relatou a existência de um *trade-off* entre as experiências psicológicas dos participantes e a eficiência da medição e preocupações de educadores(as) e especialistas quanto às especificações não estatísticas do TAI, como cobertura de conteúdo, equilíbrio de conteúdo e tamanho do teste. O autor propõe uma abordagem de teste de sombra e testes computadorizados de múltiplos estágios, para garantir a satisfação dos especialistas no assunto.

Martin e Lazendic (2018) investigaram as implicações dos testes de múltiplos estágios (MST) e testes de ordem fixa “convencionais” para vários resultados relevantes em Numeramento (Numeracia)[36], incluindo realização, motivação e engajamento relevantes para o teste e experiência de teste subjetivo. O estudo contou com 12736 participantes da escola primária australiana (anos 3 e 5) e secundária (anos 7 e 9). Realizaram modelagem multinível do nível 1 (teste) e do nível 2 (escola). As constatações confirmam que: a) testes adaptativos computadorizados proporcionam maior precisão na mensuração de resultados; b) há alguma motivação positiva relevante e efeitos de engajamento para os testes adaptativos; c) alegações contrárias de que o teste adaptativo reduz motivação, engajamento e experiência subjetiva; e (d) há efeitos positivos de testes adaptativos de computador para estudantes mais velhos em um estágio de desenvolvimento, quando eles são tipicamente menos motivados e engajados.

[36] Termos utilizados na educação matemática, equivalentes ao conceito de letramento matemático, que inserem os conhecimentos matemáticos como práticas sociais.

3.3.3.9 Estudos preocupados com ajuste do teste para alunos(as) com deficiências ou dificuldades

Os estudos em que o enfoque estava na construção de TAI para pessoas com necessidades especiais estão agrupados neste tópico.

Dolan et al. (2005) se preocuparam com a precisão das atuais avaliações de larga escala, mostrando fatores que causaram prejuízos na precisão da aferição do conhecimento de estudantes com deficiências, como as barreiras de acesso. O teste ajustado, como a leitura em voz alta, levou a melhorias, mas as descobertas da pesquisa sugerem a necessidade de uma abordagem mais flexível e individualizada dos ajustes. Um estudo piloto para a criação de um protótipo de ferramenta de entrega de TBC, que fornece aos(às) alunos(as) um ambiente flexível e personalizável tanto com a opção de ler em voz alta, quanto com o ajuste do conteúdo do teste. Dois métodos contrastantes foram usados para fornecer duas formas equivalentes de um teste de História Nacional e Civismo dos Estados Unidos para dez alunos(as) do ensino médio com dificuldades de aprendizagem. Os resultados da aplicação de duas formas equivalentes de testagem, o tradicional papel e lápis (PPT) e o TBC com a possibilidade de leitura automatizada do texto, ou texto para fala (TTS) indicaram um aumento significativo nas pontuações da administração de TBC-TTS *versus* PPT para questões com leitura de passagens com mais de 100 palavras. Os resultados qualitativos de pesquisas com os(as) alunos(as), entrevistas estruturadas, observações de campo e o rastreamento de uso realizados para obter informações sobre as preferências dos(as) alunos(as) e os padrões de uso, também apoiaram a eficácia do TBC-TTS em relação ao PPT. Os resultados deste estudo piloto forneceram suporte preliminar para os possíveis benefícios e usabilidade das tecnologias digitais na criação de testes projetados de forma mais justa e precisa para estudantes com deficiências.

Universal (2010) é um guia para ajudar os desenvolvedores de itens a identificar as diretrizes do Desenho Universal para Testes Baseados em Computador (UD-CBT). O UD-CBT forneceu uma base para o uso de tecnologias digitais na criação de testes que avaliem com maior precisão pessoas que possuem uma gama diversificada de habilidades e desafios físicos, sensoriais e cognitivos. Concentraram os estudos na compreensão da variação do escore observado no nível do item. Foram apresentadas opções de *design* que podem garantir que o conteúdo do item está alinhado com os

objetivos da medição, e também considerações sobre textos, imagens, áudios, tabelas e gráficos, notações matemáticas e científicas, vídeo e animação, ativação de objetos e *links*, opções construção de resposta (textos e matemática) e testes de múltiplos estágios.

Stone e Davey (2011) apontaram a resistência daqueles que atuam no campo da Educação Especial com relação aos testes adaptativos, mesmo com o interesse crescente no desenvolvimento de CAT e em testes de múltiplos estágios para avaliações de responsabilidade K-12. A preocupação ocorre com a forma de encaminhamento de estudantes com perfis divergentes e baixo desempenho em questões iniciais dos testes, por exemplo, alguns estudantes com dificuldades de aprendizagem em Matemática básica podem ter dificuldades com cálculos básicos, mas não com alto nível de resolução de problemas. Os autores apresentaram uma revisão de literatura com foco em questões de testes adaptativos para estudantes com deficiências no setor de ensino fundamental e médio. Concluíram que o desenvolvimento de políticas com relação a esse tópico será útil tanto para o desenvolvimento de agendas de pesquisa quanto para informar estados que estão atualmente usando ou pensando em mudar para o CAT.

Stone, Laitusis e Cook (2015) revelaram que, na última década, os avanços na tecnologia trouxeram oportunidades para melhorar a acessibilidade de avaliações para indivíduos com deficiências e, ao mesmo tempo, aumentar o envolvimento e a precisão da medição por meio de testes adaptativos. Esses avanços incluíram a integração dos princípios do *design* universal na criação de itens e a integração de tecnologias assistivas no desenvolvimento de plataformas de CBT. Além disso, os avanços nos testes adaptativos permitiram uma melhor mensuração de faixa mais ampla de níveis de desempenho, não apenas de estudantes com desenvolvimento típico. Esses avanços, no entanto, vêm com desafios adicionais na integração de conteúdos e sistemas de entrega de itens, garantindo que avaliações adaptativas fornecidas por computador não tenham consequências inesperadas para alunos(as) com deficiências. A colaboração contínua poderá fornecer avaliações justas e válidas das habilidades de todos os(as) respondentes.

Geisinger (2015) argumentou que diferentes modelos de TAI mereceram muita pesquisa. Tal pesquisa precisa comparar diferentes modelos de testes e diferentes tecnologias distintas, como *laptop, tablets* e outros dispositivos móveis para

determinar a comparabilidade das pontuações. Estudos apontaram que fazer um teste em um *laptop* com a tela e teclado menores pode deixar certos tipos de testes mais lentos ou mais difíceis. Talvez o requisito fundamental seja a segurança do teste e se os testes foram justos para membros de diferentes grupos, especialmente indivíduos com deficiência e de baixo *status* socioeconômico. As possibilidades foram vistas com entusiasmo.

Dolan e Burling (2018) esclareceram que as tecnologias digitais podem ser usadas durante os testes de várias maneiras, desde a criação de itens até o relatório dos resultados dos(as) alunos(as). Restringindo seu estudo ao uso dessa tecnologia na administração de testes a estudantes – comumente chamados de TBC – e nos usos e relatórios de dados de avaliação digital, forneceram uma visão geral do uso do TBC e algumas dicas sobre o futuro a longo prazo dos testes no ensino superior. Também mostraram que os TBC favorecem a precisão nos testes de estudantes com deficiências.

3.3.3.10 Estudos sobre o critério de parada baseados ou não na classificação dos(as) respondentes

O critério de parada, etapa essencial de um TAI, estabelece a(s) regra(s) para o encerramento do teste, buscando a minimização do erro de medida na aferição de proficiência. Essas regras são modeladas conforme o objetivo do TAI, que podem ser para a classificação do(a) respondente em termos de um ou mais pontos de corte na escala ou o alcance de um erro de medida na estimação. Os estudos inseridos neste agrupamento enfatizam essa etapa do TAI e permitem discutir o critério de parada usado no TAI desenvolvido.

Weiss (1982) mostrou as abordagens e, resumidamente, os resultados da pesquisa para testes adaptativos baseados na TRI. Testes adaptativos combinam *design* de itens e uso de diferentes regras de parada que podem ser projetadas: (1) para melhorar a qualidade e eficiência da medição, resultando em medidas de igual precisão em todos os níveis de proficiência; (2) para melhorar a eficiência da medição de baterias de teste, usando *pools* de itens projetados para o teste de administração convencional; e (3) para melhorar a precisão e eficiência dos testes para classificação (por exemplo, teste de domínio). Os resultados da pesquisa mostraram que os testes

baseados na TRI podem alcançar medidas de igual precisão em todos os níveis de proficiência, dado um conjunto de itens adequadamente projetado. Os estudos também mostraram melhorias na fidedignidade do teste e diminuição pela metade do número de itens em relação aos testes convencionais. Testes adaptativos projetados para classificação dicotômica também representam melhorias em relação aos testes convencionais projetados para o mesmo propósito. Estudos de simulação despontaram reduções no comprimento do teste e melhorias na precisão da classificação para testes adaptativos *versus* convencionais. Os dados revelaram que os testes adaptativos baseados em TRI melhoram a qualidade e/ou a eficiência da medição para cada examinando(a).

Kingsbury e Weiss (1983) compararam procedimentos para determinar o grau de proficiência com as categorias domínio e não domínio. Mostraram que medir o desempenho em relação aos objetivos de aprendizagem pré-especificados é atraente para que os(as) educadores(as): a) determinem o grau de proficiência dos(as) educandos(as) de uma sala de aula e b) usem como uma ferramenta de diagnóstico para identificar os indivíduos que necessitam de mais formação em áreas de instrução específicas. Usando simulação, observaram que, se o banco inclui itens com parâmetros *a, b* e *c* variáveis, o procedimento teste de razão de probabilidade sequencial (SPRT) resulta em testes mais curtos, mas os outros procedimentos farão classificações mais precisas. O procedimento AMT, também denominado Bayes sequencial (SB), fornece a combinação ideal e correspondência entre a alta decisão e a duração curta do teste. O procedimento AMT faz uso do intervalo de confiança e analisa quando esse intervalo não contém o ponto de corte usado para categorizar o domínio.

Spray e Reckase (1996) compararam dois procedimentos de classificação dos(as) examinandos(as) em categorias, um baseado no teste de razão de probabilidade sequencial (SPRT) e outro na metodologia de Bayes sequencial (SB), para determinar quais requeriam menos itens para classificação, quando os procedimentos foram combinados com as taxas de erro de classificação. Os resultados mostraram que, nas condições estudadas, o procedimento SPRT exigiu menos itens no teste do que o procedimento de Bayes sequencial para atingir o mesmo nível de precisão de classificação.

Eggen e Straetmans (2000) relataram que os testes adaptativos computadorizados foram desenvolvidos originalmente para obter uma estimativa eficiente da capacidade de um(a) examinando(a). No entanto, eles também se mostram úteis em problemas de classificação.

Thompson (2009) ponderou que as diversas alternativas para algoritmos de seleção de itens baseadas na TRI em testes de classificação não têm trazido evidências conclusivas sobre a superioridade substancial de uma delas. Argumentou que a falta de evidência ocorre porque os métodos avaliam os itens de maneira muito similar e geralmente selecionam o mesmo item. A consideração de métodos de seleção de itens que adotam uma faixa mais ampla do banco é frequentemente desnecessária sob condições realistas, embora possa ser vantajoso utilizá-las apenas no início de um teste. Além disso, o autor demonstrou que a eficiência das abordagens de seleção de itens depende dos critérios de finalização utilizados. A seleção de itens nos pontos de corte, que parece conceitualmente apropriada para a classificação, nem sempre é a opção mais eficiente. Uma estrutura ampla para a seleção de itens em testes de classificação foi apresentada e incorporou os pontos discutidos.

Parshall et al. (2002), Spray e Reckase (1994) e Weiss e Kingsbury (1984) descreveram CAT para situações em que o principal interesse não é estimar a capacidade de um(a) examinando(a), mas classificá-lo(a) em uma de duas categorias (por exemplo, *pass-fail* ou *master-nonmaster*). Com o estudo, os autores exploraram, via simulações, as possibilidades do CAT, com base na TRI, em classificar examinandos(as) em uma de três categorias. De modo geral, os resultados do estudo mostraram uma redução de pelo menos 22% no número médio de itens em um teste adaptativo computadorizado, comparado a um teste de papel e lápis existente. Concluiu-se que a imposição de restrições à estratégia de seleção de MFI não afeta negativamente a qualidade dos algoritmos de teste. Contudo, advertem que os profissionais devem ser cautelosos em generalizar desses resultados para aplicações que usam procedimentos de controle de exposição e algoritmos complexos de balanceamento de conteúdo.

Babcock e Weiss (2012) utilizaram simulação para investigar vários critérios de parada para os TAI. No estudo, utilizaram várias regras básicas de parada: erro padrão SE máximo, mudanças na proficiência (θ), comprimento fixo do teste e combinações delas. Essas combinações foram usadas para aproveitar tanto a alta precisão de

medição quanto a possibilidade de terminar rapidamente o teste quando a alta precisão não era possível. Os métodos foram verificados para quatro bancos de itens diferentes. Várias conclusões foram encontradas: a) que, ao contrário das afirmações na literatura, os TAI de comprimento variável não foram enviesados nem tiveram desempenho pior do que os de comprimento fixo; b) que os TAI de comprimento variável tiveram um desempenho igual ou ligeiramente melhor do que os seus homólogos de comprimento fixo, quando os comprimentos médios dos testes eram comparáveis; c) que resultados anteriores, alegando que TAI de comprimento variável são enviesados, são devidos a técnicas de estimação bayesianas combinadas com TAI de comprimento variável que terminaram com poucos itens; d) que a melhor solução para o critério de parada foi usar um ou mais critérios de comprimento variável combinados com um número mínimo de 15 a 20 itens, dependendo das necessidades de precisão do usuário do teste; e) um critério de parada de comprimento variável suplementaria o critério mínimo de itens, administrando mais itens a pessoas que ainda não foram bem medidas; f) o critério de informações máxima, como uma regra suplementar para o de erro padrão, também é uma alternativa viável para um banco de itens pequeno; g) a pesquisa tem limitações, pois não controlou a exposição de itens ou o balanceamento de conteúdo.

Spenassato et al. (2015) apresentaram as vantagens dos TAC em relação aos métodos não adaptativos aplicados via computador ou papel e lápis, para selecionar itens mais informativos para os(as) respondentes. O trabalho focou em realizar a avaliação da maturidade do Sistema de Gestão Ambiental (SGA) das indústrias. O teste aborda questões sobre política ambiental, planejamento, implementação e operação, verificação e ação corretiva entre outras. Via simulação, foi estudado o efeito de diferentes regras de finalização do TAC e os resultados foram comparados à sua versão original. Concluiu-se que um TAC poderia reduzir em 71% o tamanho do teste, sem comprometer a validade da medida.

3.3.4 Conclusões acerca da revisão da literatura sobre o TAI

Com os avanços da tecnologia, os testes para avaliação educacional sofreram mudanças significativas. As primeiras mudanças tecnológicas aconteceram pela utilização de *softwares* e computadores na análise das respostas aos itens, permitindo

o uso de modelos avançados de análise estatística, como a TRI, nos processos psicométricos. Os mais recentes avanços têm acontecido pelo uso dos computadores na construção e na aplicação dos testes, embora essa utilização seja ainda pontual no Brasil. Para a construção de um TAI, a literatura apontava como básico: 1) um banco de itens calibrados e 2) um algoritmo para selecionar item, aferir a proficiência e encerrar o teste.

O algoritmo desenvolvido para o TAI experimental da PB – Leitura abarcou uma regra adicional no critério de parada, que considera a classificação da proficiência, analisando o intervalo de confiança. Os TAI usualmente têm se voltado para estimar a proficiência de modo mais preciso (medida de um atributo), sem se preocupar em classificar essa proficiência em determinado nível de interpretação pedagógica da escala. A classificação (Cf. BABCOCK; WEISS, 2012; KINGSBURY; WEISS, 1983; SPRAY; RECKASE, 1994, 1996; WEISS, 1982; WEISS; KINGSBURY, 1984), mais usada em situações de avaliação somativas, nas quais uma decisão do tipo sim/não era tomada, é mais comum em exames e provas de certificação.

O aspecto diferencial do critério de parada do TAI da PB – Leitura desenvolvido se apoiou nas características presentes no teste em papel e lápis da PB – Leitura, totalmente voltado para o diagnóstico da proficiência em leitura nos anos iniciais da escolarização e na alocação da proficiência do(a) aluno(a) em um dos cinco níveis da escala para o construto da leitura, fruto do processo de medida educacional determinado por um painel de professores(as) que agregou validade pedagógica. O teste da PB – Leitura em papel e lápis e seu correspondente em meio digital compreendeu um conteúdo que mantém total identidade com as ações pedagógicas cotidianas de professores(as) alfabetizadores(as), permitindo que os resultados do teste fossem usados na avaliação de caráter formativo. Assim, a classificação do(a) aluno(a) em um dos níveis da escala está diretamente relacionada à avaliação formativa da competência leitora e, por sua vez, pode pressupor intervenção educativa real e direta por parte do(a) professor(a).

Dessa forma, o critério de encerramento agregou uma regra usada em teste com objetivo de classificação, que procurava verificar a existência de um dos pontos de corte da escala no intervalo construído a partir da proficiência estimada. Embora o objetivo se volte para classificação, a acepção de classificação está vinculada ao aspecto de categorização, mais avançado do que a ação de medir, que possibilita a

atribuição de números e a possibilidade de ordenar. Essa acepção se volta para a possibilidade de diferenciar características de domínios que possibilitam intervenções pedagógicas diferenciadas. Na classificação, não se abandona a ideia de ordenação, mas se incorpora a ideia de classe, *clusters*.

Foram combinadas regras que não visavam unicamente a medida precisa da proficiência, que prevê diminuição e homogeneização da dimensão do erro da medida, mas também em alocar completamente o intervalo de confiança da proficiência em um dos cinco níveis da escala, aspecto que evoca a ideia de que nem sempre a medida precisa é essencial para identificar em qual nível se situa o domínio do(a) respondente e, consequentemente, as intervenções pedagógicas necessárias.

Esse último aspecto passou a ter relevância na construção do TAI da PB – Leitura, ao possibilitar que, mesmo quando o erro de medida não atingia patamares aceitáveis, seria possível atribuir precisão na definição do nível de proficiência, servindo para enfatizar o caráter formativo da avaliação.

4 O PROCESSO DE CONSTRUÇÃO DO TAI DA PB – LEITURA

Neste capítulo é apresentada a construção do TAI da PB – Leitura, descrevendo as etapas necessárias para isso a partir de um teste já existente em papel e lápis da PB – Leitura.

Essa situação se colocou mediante algumas facilidades, como a existência de um conjunto de elementos que caracterizavam a PB como uma avaliação, aspecto salientado em capítulo anterior, inclusive com ampla utilização, até 2016, por professores(as) alfabetizadores(as) e a existência de um banco de itens parametrizado, formado pelos itens aplicados em edições anteriores da PB – Leitura. No entanto, um desafio adicional se colocou: verificar se os parâmetros dos itens do banco poderiam ser mantidos no TAI, visto serem administrados em meio eletrônico diverso do impresso. Isso se avultou para que pudéssemos garantir a mesma interpretação pedagógica dos resultados quando da alteração de mídia de administração do teste.

Fazem parte deste capítulo, seções que descrevem: o processo de articulação das escolas que participaram do Projeto; o processo de formação de suas equipes; a aplicação do TBC da PB – Leitura, incluindo o tratamento dos resultados de aplicação dos itens na versão impressa e na digital; e a aplicação do TAI da PB – Leitura, compreendendo as características do algoritmo, da escala de proficiência e a análise dos resultados.

4.1 O envolvimento de estudantes, professores(as) e gestores(as) na pesquisa

A construção do TAI da PB – Leitura foi planejada para ocorrer mediante um processo de aplicação experimental que viabilizava a obtenção de dados de alunos reais, pertencentes a escolas, com equipes docentes e gestoras igualmente reais.

Nessa perspectiva, a construção da ferramenta atenderia à expectativa de refletir sobre patamares adequados para a implementação de processos de avaliação para redes de ensino, especialmente quanto ao caráter formativo ou para disseminação desses processos no interior de escolas e salas de aula, associados, inclusive, à formação de professores(as) em avaliação da aprendizagem.

Direcionado à solução de um problema real e que se origina no processo de alfabetização, o projeto TAI para PB – Leitura configurou uma pesquisa aplicada e participativa, dada a preocupação em envolver especialistas, professores(as) e gestores(as) durante o desenvolvimento da ferramenta, significando o enfrentamento do baixo reconhecimento por parte de professores(as) e gestores(as) quanto às potencialidades das avaliações externas, sendo, em muitas situações, mais evidente concentrarem-se no reconhecimento de suas limitações, dificultando a incorporação de seus resultados nos processos pedagógicos em sala de aula ou no planejamento escolar de unidades educacionais.

O objetivo foi de verificar se é possível suprir lacunas inerentes aos processos avaliativos na alfabetização e no letramento inicial com a produção de uma versão adaptada e informatizada da PB – Leitura, envolvendo a participação de professores(as) e gestores(as) de 15 Emef[37] da RME/SP, mesmo sabendo-se que esse processo pudesse ter características de um ensaio com vistas ao desenvolvimento de um TAI.

Assim, para lograr responder à questão: “É possível construir um TAI para a versão impressa da PB – Leitura que seja ponto de apoio para professores(as) na avaliação de alunos(as) dos anos iniciais do ensino fundamental?”, foi necessário estabelecer uma série de objetivos específicos, como buscar referências bibliográficas sobre os testes adaptativos informatizados; estudar as características psicométricas do teste impresso da PB – Leitura e verificar diferenças psicométricas da aplicação de uma versão informatizada dela; estudar aspectos do algoritmo de um TAI; analisar a construção e aplicação do TAI da PB – Leitura e discutir com professores(as) os pressupostos da medida educacional e a interpretação pedagógica dos resultados.

Diferente de outros estudos que envolveram construção de TAI, a pesquisa não se restringiu ao desenvolvimento de uma aplicação ou um *software*, considerando as etapas relativas ao trabalho dos profissionais das áreas de Psicometria e Engenharia informática, mas envolveu, além dos(as) alunos(as) nas aplicações dos testes, gestores(as) e professores(as) das escolas e gestores(as) da SME/SP. Essa realização aproximou a construção da ferramenta dos principais atores da prática educativa, de forma que gestores(as), professores(as) e alunos(as) não foram

[37] Escola Municipal de Ensino Fundamental.

envolvidos apenas para responderem a questões e testarem os dispositivos. Almejou-se um processo de formação em avaliação e a constituição de uma análise crítica da interpretação pedagógica da escala.

Por um lado, em razão da inserção da pesquisa no âmbito do projeto desenvolvido pela SME/SP, por intermédio do Núcleo Técnico de Avaliação (NTA), e a decisão por abranger de maneira participativa escolas e educadores(as), foram envolvidos aspectos da realidade que não podiam ser quantificados e cuja compreensão estava centrada na explicação das interações sociais, de natureza qualitativa. Por outro, a coleta de dados sobre as aplicações das versões TBC e TAI da PB – Leitura envolveu a abordagem quantitativa das informações.

Somente integrada ao Projeto TAI da PB – Leitura, que envolveu escolas, pesquisadores da Feusp, especialistas do Inep e da SME/SP, seria possível uma pesquisa com as características em tela. Esse intento permitiu a articulação de engenheiros de *softwares*, elaboradores de testes, pesquisadores e profissionais da educação, imprescindível ao desenvolvimento de um TAI, como enfatizam Yan, von Davier e Lewis (2014).

A SME/SP subsidiou a contratação de assessor para desenvolver a plataforma informatizada, responsável pela exibição de itens aos(às) respondentes e pela captura automática das respostas, e de assessor para programar o algoritmo do TAI. Também foi responsável pela disponibilização de *tablets* e de estagiários(as). O Inep disponibilizou para SME/SP os relatórios técnicos sobre a realização dos pré-testes de itens e os parâmetros psicométricos dos itens dos testes 1 e 2 da edição de 2015 da PB – Leitura.

Houve um levantamento bibliográfico sobre a construção do TAI que possibilitou identificar as vantagens e limitações na sua implementação. A abordagem quantitativa foi utilizada, sobretudo no momento das análises das aplicações da versão eletrônica da PB – Leitura e na análise dos resultados do TAI da PB – Leitura, aplicados para os(as) alunos(as) do 1º e 2º anos do ensino fundamental das 15 escolas do estudo de campo.

A abordagem quantitativa não significou a adoção de processo rigoroso de amostragem de estudantes, visto que essa adoção significaria uma perda quanto ao envolvimento espontâneo das equipes das escolas, crucial para o processo de

pesquisa participativa, prejudicando os momentos de reflexão sobre avaliação, matriz de avaliação e sua relação com a alfabetização e letramento inicial, aspectos essenciais ao desenvolvimento da pesquisa. Assim sendo, considerou-se o estudo como censitário voluntário, no que diz respeito às escolas e respectivos(as) educadores(as) envolvidos na pesquisa.

Em paralelo ao levantamento bibliográfico e às aplicações dos testes eletrônicos e adaptativos, tendo em vista que a pesquisa envolveu um processo de formação de professores(as) sobre avaliação da alfabetização e entendido como um processo de investigação, de educação e de ação, envolvendo professores(as) de 15 unidades educacionais, articulando "[...] o processo de geração e o de uso do conhecimento, entre o mundo 'acadêmico' e o 'real', entre intelectuais e trabalhadores, entre 'ciência' e 'vida'"(HAGUETTE, 2014, p. 147), caracterizou-se a pesquisa como participante (Cf. GERHARDT; SILVEIRA, 2009). Além disso, a construção do TAI a partir de um teste preexistente e com finalidade significativa para a educação brasileira, de caráter formativo, estabeleceu diferencial em relação a outras pesquisas brasileiras que abrangeram a constituição de um TAI.

A proposta de materialização do TAI da PB – Leitura, contando com a participação de educandos(as), professores(as) e gestores(as), além dos(as) pesquisadores(as), seria importante para trazer novos horizontes na filosofia e nas práticas avaliativas.

O projeto abarcou unidades escolares de duas das treze Diretorias Regionais de Educação (DRE) que integram a SME/SP, tendo em vista o envolvimento de dois pesquisadores do Gepave que estavam lotados como supervisores escolares nessas duas DRE[38] e que poderiam conciliar melhor seu trabalho com a realização das atividades de pesquisa. O NTA foi decisivo para a articulação com os Diretores Regionais de Educação e equipe de supervisores escolares dessas DRE, estabelecendo o contato e apresentando argumentos favoráveis ao desenvolvimento

[38] A SME/SP congrega 13 (treze) Diretorias Regionais de Educação, órgãos distribuídos nas regiões do município de São Paulo para dar apoio administrativo e pedagógico às unidades que integram a rede educacional. Em junho de 2016, faziam parte da rede pública municipal: 8 (oito) unidades de Ensino Fundamental e Médio, as EMEFM, 45 (quarenta e cinco) unidades de Ensino Fundamental localizadas nos Centros Educacionais Unificados, denominadas CEU Emef, 501 (quinhentas e uma) unidades de Ensino Fundamental, as Emef, e 8 (oito) unidades de Educação Básica para surdos, as Emebs, além das unidades da Educação Infantil e de Educação de Jovens e Adultos.

do projeto. Esse contato foi realizado em meados de abril de 2016, ano em que o projeto foi desenvolvido.

A observação do perfil de matrículas das unidades educacionais dessas DRE serviu para verificar que 15 unidades seriam suficientes para garantir um número de estudantes matriculados(as) no 2º ano do ensino fundamental, população-alvo da PB, atendendo à quantidade mínima de 500 respondentes para obtenção de parâmetros, restrição estatística observada na literatura. Esse cuidado foi tomado, caso fosse necessário recalibrar os itens da PB, aspecto que não se confirmou após a análise do comportamento diferencial dos itens para os grupos que participaram da administração impressa e eletrônica (TBC da PB – Leitura), tratado na subseção 4.3.3.

Embora o delineamento inicial tenha sido para os(as) alunos(as) do 2º ano, a aplicação do TAI da PB – Leitura, dado que ocorreu no final do ano letivo, foi oportunizada também para estudantes do 1º ano, tendo em vista que esses(as) estudantes realizariam a PB no início do ano letivo seguinte, caso ela tivesse sido mantida pelo Inep.

Na reunião de esclarecimento e introdução do projeto, tanto para supervisores escolares como para os(as) diretores(as) das escolas, foram especificados os pressupostos, os objetivos, a metodologia, as fases, o cronograma, os envolvidos e as condições materiais do projeto. A adesão do(a) diretor(a) da escola e sua implicação real no projeto foram de suma importância, dado seu papel fundamental no posterior engajamento da equipe docente. Foi oferecida a possibilidade de a equipe gestora da escola optar por aderir ao projeto após consulta aos(às) professores(as) do ciclo de alfabetização[39]. Essa prerrogativa seria decisiva para que o projeto não fosse compreendido como uma decisão unicamente externa, garantindo a característica de pesquisa participante.

Foram enumeradas nove fases do Projeto, que consistiram de:

1ª. Reunião de esclarecimento e introdução ao TAI da PB – Leitura, realizada de abril a junho de 2016;

[39] Na SME/SP, os professores regentes do 1º, 2º e 3º anos do ensino fundamental atuam no denominado ciclo de alfabetização.

2ª. Minicurso de introdução à medida educacional, realizada de abril a junho de 2016;

3ª. Realização de visitas técnicas de verificação das condições tecnológicas para aplicação da prova informatizada e do TAI, realizada de agosto a novembro de 2016;

4ª. Aplicação da versão informatizada da PB – Leitura, realizada de setembro a outubro de 2016;

5ª. Discussão sobre a Matriz de referência e a intepretação pedagógica da PB – Leitura com professores(as), realizada de agosto a setembro de 2016;

6ª. Aplicação do protótipo adaptativo informatizado da PB – Leitura, realizada em outubro de 2016;

7ª. Aplicação do TAI da PB – Leitura, realizada em novembro de 2016;

8ª. Realização de seminário sobre alfabetização, atividade prevista para o final do projeto;

9ª. Realização de seminário sobre TAI, em dezembro de 2016.

O seminário de alfabetização, previsto na 8ª fase, acabou sendo prejudicado devido ao atraso nas aplicações das fases anteriores, que dependiam do funcionamento dos *tablets* e das redes nas unidades. Em 2017, o projeto sofreu solução de continuidade devido à nomeação de outros profissionais pelo prefeito eleito em 2016.

O desafio seguinte foi quanto à definição de quais unidades seriam selecionadas, uma vez que a DRE 1 tinha, sob sua jurisdição, 51 unidades educacionais que disponibilizam ensino fundamental, entre Emef, EMEFM[40] e CEU Emef[41] e a DRE 2 tinha 38 unidades entre Emef e CEU Emef.

Posteriormente à manifestação das escolas interessadas, foram escolhidas 8 unidades de uma DRE e 7 unidades na DRE menor, com base na análise nos resultados de aplicações da PB – Leitura nas edições de 2013 a 2015. A escolha foi pautada nos testes aplicados no início do ano, denominado teste 1 da PB – Leitura, uma vez que faltavam os dados de 2015, da série histórica referente ao teste 2. A

[40] Escola Municipal de Ensino Fundamental e Médio.
[41] Escola Municipal de Ensino Fundamental localizada no Centro Educacional Unificado.

aplicação do teste 2, da edição de 2015 da PB – Leitura tinha sido adiada para o início de 2016 e os dados não tinham sido digitados ainda.

Assim, a partir da série histórica dos resultados do teste 1 da PB – Leitura, nos anos de 2013, 2014 e 2015, especificamente agrupados os percentuais de alunos(as) nos dois primeiros níveis, 1 e 2, e nos dois últimos níveis, 4 e 5, foram escolhidas unidades cuja série apresentava diferentes variações e dimensões. Procurou-se cobrir unidades cujos percentuais aumentavam e decaíam bem como unidades com altos e baixos percentuais de crianças nos níveis iniciais ou nos finais da PB – Leitura. Essa preocupação garantiria uma amplitude no espectro das proficiências dos(as) alunos(as) das escolas envolvidas no projeto.

Além das variações nas séries históricas e a diferenciação de percentuais nos níveis inicial e final, outros fatores foram relevantes para a seleção, por exemplo, o acesso por meio de transporte coletivo à unidade escolar, tendo em vista o deslocamento dos estagiários(as) que auxiliariam o desenvolvimento da pesquisa. Outro fator eventualmente considerado foi a existência de uma equipe técnica completa na escola (diretor, dois assistentes de diretor e dois coordenadores pedagógicos), possibilitando o acompanhamento das formações envolvidas no projeto.

Na DRE 1, das 15 Emef que manifestaram interesse em participar do projeto TAI da PB – Leitura, as oito unidades selecionadas mediante os critérios anteriormente citados são apresentadas na Tabela 1. Também se encontram nessa tabela as respectivas séries históricas em relação à distribuição dos(as) alunos(as) nos níveis da PB – Leitura das unidades educacionais selecionadas, sendo que foi suprimido o percentual de crianças no nível 3. A identificação das unidades foi trocada para preservar a privacidade das instituições escolares envolvidas. Foram escolhidos nomes de países do continente africano, dado que o Gepave desenvolvia projeto também nesse continente e decidiu prestar essa homenagem.

O cuidado se justifica diante do envolvimento de instituições e pessoas. Serão adotados os princípios das declarações e convenções sobre Direitos Humanos, expressos na Constituição Federal, no Código de Ética de Pesquisa (CEP) e na Resolução do Conselho Nacional de Saúde (CNS) nº 196/96. Sendo assim, a confidencialidade das informações, a privacidade dos sujeitos e o anonimato dos participantes e instituições serão preservados.

Cabe também observar que os dados obtidos serão utilizados exclusivamente para os fins previstos nos termos de consentimento devidamente solicitados e o apoio dos colaboradores na elaboração do trabalho e o auxílio de pesquisadores da área serão citados e reconhecidos.

Tabela 1 – Distribuição percentual de estudantes da DRE 1, por nível de desempenho no teste 1 da PB – Leitura, por agrupamento dos níveis 1 e 2 – 4 e 5

Emef	2013		2014		2015	
	Níveis 1 e 2	Níveis 4 e 5	Níveis 1 e 2	Níveis 4 e 5	Níveis 1 e 2	Níveis 4 e 5
Camarões	1,72	93,97	18,75	46,09	7,41	54,63
Ruanda	4,21	89,47	8,70	55,65	13,16	56,14
África do Sul	3,92	89,22	29,69	39,06	15,38	48,08
Moçambique	5,19	79,22	18,89	42,22	17,72	48,10
Marrocos	0,00	100,00	20,63	25,40	18,18	29,09
Benim	2,52	97,48	23,44	31,25	21,31	32,79
Egito	n/c	n/c	28,13	12,50	23,08	15,38
Libéria	5,13	92,31	27,42	27,42	25,42	28,81

*n/c (não consta) porque a escola não havia sido criada.
Fonte: Elaboração da autora, com base nas informações do Núcleo Técnico de Avaliação da SME/SP.

É possível notar diferenças importantes entre as unidades, quanto à dimensão e variação nos percentuais de crianças: as unidades definidas apresentam variação ora crescente, ora decrescente nos percentuais de 2013 a 2015, e representam unidades com percentuais de crianças nos níveis 1 e 2 menores ou iguais a 15,5% (3 unidades em 2015) e percentuais acima de 15,5% (5 unidades em 2015).

A Tabela 2 evidencia os quantitativos para os(as) estudantes matriculados no 2º ano das escolas da DRE 1 com os quais serão realizadas as aplicações informatizadas.

Tabela 2 – Quantitativos referentes ao 2º ano do ensino fundamental nas escolas da DRE 1, por turmas e por estudantes (data-base 19/07/2016)

Emef	2016	
	nº de turmas do 2º ano	nº de estudantes de 2º ano
Camarões	3	100
Ruanda	3	94
África do Sul	2	67
Moçambique	3	91
Marrocos	3	96
Benim	2	60
Egito	3	97
Libéria	3	90
Total	22	695

Fonte: Elaboração da autora, com base nas informações do Sistema Escola On-Line (EOL) da SME/SP.

Na DRE 2, 10 unidades escolares se inscreveram para participar do projeto TAI e 7 unidades foram selecionadas mediante os critérios citados. A Tabela 3 apresenta as séries históricas das unidades educacionais selecionadas e destaca-se, mais uma vez, que não são apresentados os percentuais referentes ao nível 3.

Tabela 3 – Distribuição percentual de estudantes da DRE 2, por nível de desempenho no teste 1 da PB – Leitura, por agrupamento dos níveis 1 e 2 – 4 e 5

Emef	2013		2014		2015	
	Níveis 1 e 2	Níveis 4 e 5	Níveis 1 e 2	Níveis 4 e 5	Níveis 1 e 2	Níveis 4 e 5
Líbia	8,33	87,50	16,13	35,48	10,71	39,29
Angola	1,32	96,05	5,56	81,11	9,26	67,59
Cabo Verde	1,20	95,18	10,23	50,00	10,47	51,16
Etiópia	6,56	86,89	24,71	35,29	12,86	42,86
República do Congo	19,63	78,50	18,75	38,28	16,81	43,36
Costa do Marfim	1,33	89,33	17,86	35,71	22,62	35,71
Argélia	0,00	97,83	20,83	33,33	25,00	36,36

Fonte: Dados organizados pela autora, com base nas informações do Núcleo Técnico de Avaliação da SME/SP.

Nessa DRE, também foram definidas unidades que além de apresentarem diferenças de variação, na série de 2015, apresentavam percentuais de crianças nos níveis 1 e 2 menores ou iguais a 15% (4 unidades) e percentuais acima de 15% (3 unidades).

A Tabela 4 apresenta o número de estudantes matriculados nas turmas do 2º ano das escolas da DRE 2, envolvidos nas aplicações informatizadas.

Tabela 4 – Quantitativos referentes ao 2º ano do ensino fundamental nas escolas da DRE 2, por turmas e por estudantes (data-base 19/07/2016)

Emef	2016	
	nº de turmas do 2º ano	nº de estudantes de 2º ano
Líbia	2	46
Angola	3	87
Cabo Verde	3	91
Etiópia	3	90
República do Congo	4	126
Costa do Marfim	3	83
Argélia	2	60
Total	20	583

Fonte: Dados organizados pela autora, com base nas informações do Sistema Escola On-Line (EOL) da SME/SP.

Cerca de 80 professores(as) e 30 gestores(as) das unidades, que atuavam nos anos iniciais do ensino fundamental, participaram de reuniões para formação específica em medida educacional, cujo objetivo foi estabelecer um patamar mínimo de informações

sobre aspectos da medida educacional, e de uma reunião de esclarecimento sobre as características do projeto de pesquisa e introdução ao TAI da PB – Leitura, ambas realizadas de abril a julho de 2016. Sobre esse assunto, destina-se a seção a seguir.

4.2 O trabalho de formação de professores(as) alfabetizadores(as)

Um pressuposto do estudo era o de que a produção do dispositivo eletrônico para aplicação de provas adaptadas poderia ser espaço de reflexão sobre o processo de avaliação da proficiência em leitura, sem reduzi-lo ao ato de transferir para uma máquina os desafios de avaliar. Nessa perspectiva, o projeto de pesquisa visou à construção de um artefato avaliativo que tivesse relevância para o processo de ensino e aprendizagem, essência da avaliação. Por esse motivo, a formação docente em avaliação da proficiência em leitura nos anos iniciais do ensino fundamental, realizada no contexto do desenvolvimento do TAI, possibilitou a formação de cerca de 80 professores(as) em um processo de formação que priorizou o estudo de elementos da avaliação formativa e da medida educacional.

Conforme assinalaram Alavarse et al. (2018a, 2018b), a formação dos professores(as) foi balizada pelas considerações de que:

a) a alfabetização é um dos principais objetos de ensino nos anos iniciais do Ensino Fundamental e um dos objetivos mais relevantes da educação no mundo inteiro, visando à democratização do acesso a uma das mais importantes competências de nossa cultura, como destacou Williams (2000);
b) a plena alfabetização está longe de ser atingida, haja vista as taxas de reprovação nos anos iniciais, fruto de decisões de professores(as) reconhecidamente focadas nas competências leitora e escritora, e os resultados de avaliações externas em larga escala, reveladores de baixas proficiências para a imensa maioria dos(as) alunos(as) dos anos iniciais;
c) há dificuldade em avaliar a alfabetização na sala de aula, especialmente quando se considera o conceito de avaliação dado por Luka Mujika e Santiago Etxebarría (2009), que a concebem como julgamento de algo visando à sua melhoria, ou por Nevo (1998), que considera a necessidade do diálogo entre avaliações externa e interna;

d) a PB – Leitura é um teste padronizado para diagnóstico, amplamente utilizado no território nacional, na qual se encontram elementos característicos de uma avaliação com fundamentos documentados;
e) há um reconhecimento, de um lado, que os riscos em se tratar a alfabetização de modo mecânico e acrítico, sem considerar o contexto e as expectativas dos(as) alunos(as), conforme adverte Street (1993), poderia levar ao que se denomina modelo autônomo de alfabetização, de outro, que o rompimento com a seletividade e as práticas fragmentadas na alfabetização não é um objetivo que se resolva no campo da avaliação, mas nele aparecem os elementos que podem incorporar estratégias para processos pedagógicos inclusivos.

Essa perspectiva permeou a formação com professores(as) e gestores(as) das 15 Emef e demandou a elaboração de uma agenda que possibilitasse a articulação entre os horários coletivos de trabalho pedagógico das 15 escolas e os horários dos formadores envolvidos no projeto. Essa agenda tem relevância porque a formação foi realizada na própria unidade escolar, dentro do horário institucionalizado para essa ação, considerando a presença de professores(as) alfabetizadores(as) e gestores(as) de cada uma dessas unidades escolares.

Os(as) pesquisadores(as)/formadores(as) do Gepave e do NTA foram coordenados(as) pelo Dr. Ocimar Alavarse, no processo de formação realizado na segunda fase, que promoveu o minicurso de introdução à medida educacional, e pelo professor Ailton Carlos Santos, no processo de formação realizado na quinta fase. Foi organizada uma logística de atendimento às unidades, compondo uma agenda de três encontros para cada um dos focos de estudo apresentados a seguir e para cada uma das 15 escolas:

1) introdução às medidas educacionais: suas possibilidades e limitações;

2) fundamentos teórico-metodológicos da avaliação realizada por meio da PB – Leitura.

Os três encontros, referentes ao foco 1, constituíram a primeira fase de formação de educadores(as) sobre a construção do TAI, problematizando e aproximando os professores(as) e gestores(as) da análise psicométrica de itens de provas padronizadas que tinham sido aplicadas em todas as unidades da rede (cerca de 600 unidades), mostrando os limites e as possibilidades de uma medida educacional.

Os três encontros, referentes ao foco 2, constituíram a quinta fase da construção do TAI da PB – Leitura e proporcionaram reflexões sobre:

a) Concepções de alfabetização e letramento inicial e suas relações com a PB – Leitura;

b) Matriz de Referência para Alfabetização e Letramento Inicial da PB – Leitura, suas relações com o currículo da alfabetização e perspectivas para um TAI;

c) Interpretação de resultados nos documentos da PB: relações entre os níveis de desempenho, as habilidades da matriz de referência e as intervenções pedagógicas.

As reflexões realizadas nos encontros foram anotadas pelos formadores e discutidas em reuniões centrais. Nesta seção, serão descritas as reflexões relativas aos encontros da quinta fase, pois as reflexões referentes à primeira fase foram foco de outra investigação realizada no âmbito do Gepave.

4.2.1 Dos encontros sobre fundamentos teórico-metodológicos da PB – Leitura

Um pressuposto central na formação implementada com a construção do TAI era o de que a avaliação fosse vista como parte do processo de ensino e de aprendizagem e considerada no seu aspecto formativo, possibilitando detectar o nível de alfabetização dos(as) alunos(as) e possíveis insuficiências nas habilidades de leitura e escrita no 2º ano do ensino fundamental.

Para que a avaliação tivesse essa perspectiva, seria necessário discutir as bases epistemológicas que fundamentam a construção do teste da PB – Leitura na formação. Assim, o primeiro encontro formativo foi destinado a identificar os desafios a enfrentar no processo de alfabetização e letramento inicial, refletir sobre as concepções correntes de professores(as) sobre esse processo e suas relações com o teste da PB – Leitura.

As discussões colocaram em pauta elementos fundamentais para reflexão sobre a prática docente e revelaram que, embora os(as) professores(as) conhecessem os métodos de alfabetização denominados tradicionais e a abordagem construtivista, permaneciam os desafios apresentados no processo de desenvolvimento da proficiência em leitura e aquisição da escrita, relacionados ao que Soares (2016) denominou “alfabetizar letrando”.

Igualmente buscou-se debater nesse encontro o que Micotti (2013) apontou como enfoque interativo dos vários aspectos do desenvolvimento da leitura e da aquisição da escrita, propondo uma prática pedagógica que considerasse a abordagem de vários tipos de textos em situações reais de comunicação, a fim de desenvolver a compreensão e os procedimentos cognitivos que entram em jogo nessa aquisição, sem perder de vista o trabalho com os elementos constitutivos da escrita em suas relações com a língua oral, ou seja, a leitura.

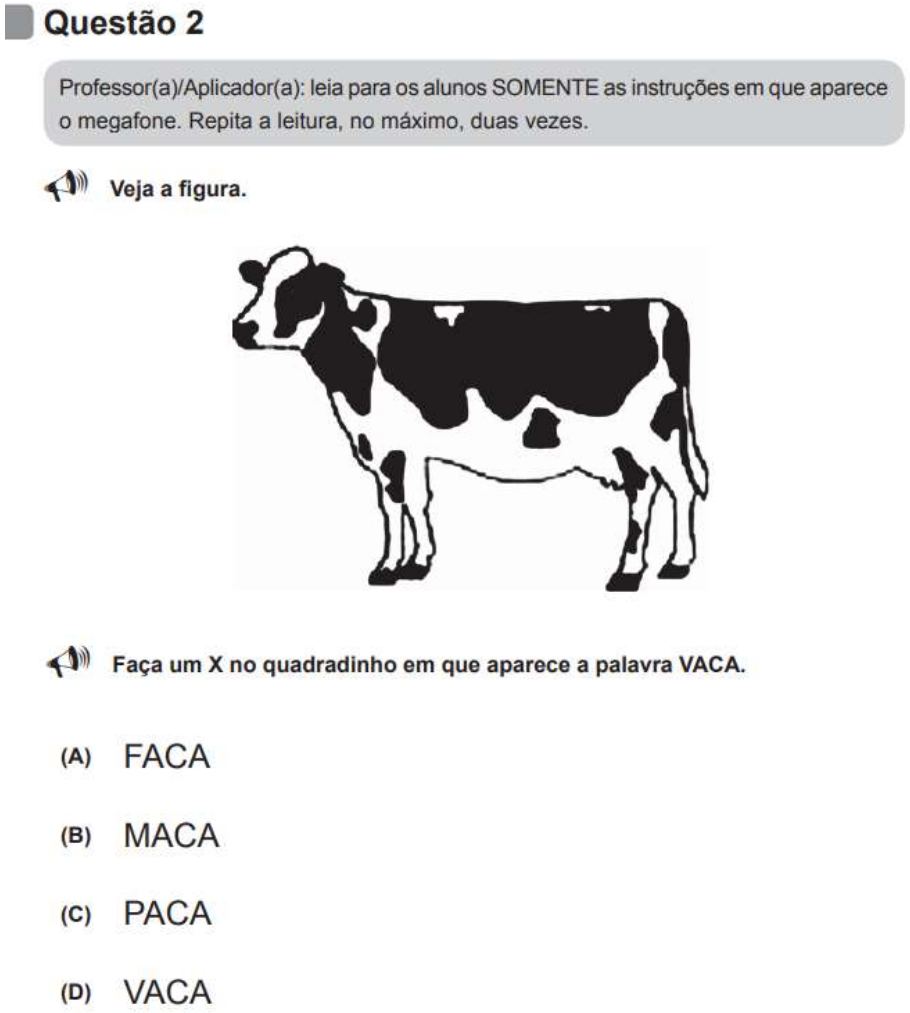
O debate sobre esses elementos, no primeiro encontro, ocorreu fundamentalmente quando os(as) professores(as) ser referiam aos textos que constituíram os itens da PB – Leitura. Esses textos foram considerados muito difíceis para alunos(as) do processo de alfabetização e letramento inicial, aspecto que colidia com as afirmações das autoras supracitadas, tendo em vista que elas defendem uma abordagem sobre a alfabetização na perspectiva do letramento. Essa perspectiva considera que a leitura deve abranger desde capacidades necessárias ao processo de alfabetização até aquelas que habilitam o(a) aluno(a) à participação ativa nas práticas sociais letradas, aquelas que contribuem para o seu letramento (BRASIL, 2008, p. 39). Nesse contexto, os textos considerados difíceis pelos(as) professores(as) envolvidos(as) na formação faziam parte de situações reais de leitura em que os(as) alunos(as) estão implicados, ainda que não apresentassem o domínio pleno da leitura.

Além disso, foi observada uma preocupação dos(as) professores(as) envolvidos(as) com a consciência fonológica, visto que há ocorrência de itens que exigiam a habilidade de relacionar as unidades sonoras às suas representações gráficas nos testes da PB – Leitura. Sobre esse assunto, Morais (2012, p. 84) observou a existência de relativo consenso sobre o que é chamado de “consciência fonológica” como uma grande “constelação” de habilidades, envolvendo a reflexão sobre os segmentos sonoros das palavras, de tal modo que:

> A consciência fonológica não é uma coisa que se tem ou não, mas um conjunto de habilidades que varia consideravelmente. Uma primeira fonte de variação é o tipo de operação cognitiva que fazemos sobre as partes das palavras: pronunciá-las, separando-as em voz alta; juntar partes que escutamos separadas; contar as partes das palavras; comparar palavras quanto ao tamanho ou identificar semelhanças entre alguns pedaços sonoros; dizer palavras parecidas quanto a algum segmento sonoro etc.

As discussões realizadas com os(as) professores(as) no processo de formação evidenciaram a necessidade de trabalhar as habilidades de consciência fonológica de forma intencional e sistemática pois, “compreender como se processam as relações entre grafemas e fonemas é fundamental para o processo de aprendizagem do funcionamento do sistema de escrita alfabética” (MORAIS 2012, p. 84). Certos itens da PB – Leitura, cujas habilidades envolviam as unidades sonoras e suas representações gráficas, no entendimento dos(as) professores(as), exigiam a consciência fonêmica, isto é, uma relação direta letra-som, quando na verdade, a reflexão sobre os segmentos sonoros das palavras envolvia aspectos mais complexos inerentes à consciência fonológica, visto que consideravam de forma contextualizada a relação letra-som, como no exemplo da Figura 4.

Figura 4 – Questão da PB – Leitura, teste 2, edição 2015



Fonte: Brasil (2015b).

No item da Figura 4, que é a versão ilustrada no guia do(a) professor(a)/aplicador(a) e por isso não mostra os quadradinhos que serão visualizados pelo(a) aluno(a), dentre as possibilidades de variações, destaca-se a necessidade de identificar que as palavras nas alternativas terminam com som parecido, mas o som inicial é que as diferencia, não só porque são grafadas de forma diferente, mas porque o sentido varia de acordo com a mudança sonora. Desse modo, ser capaz de identificar palavras que compartilham o mesmo fonema (e não toda a sílaba) inicial é importante para se alcançar uma hipótese silábico-alfabética ou alfabética da escrita (MORAIS, 2012, p. 87). Saber apenas que houve uma mudança de letra ou da representação gráfica do som não é suficiente para que o(a) aluno(a), por meio da leitura, adquira consciência de que a unidade sonora é provocadora de sentido da palavra como um todo.

É notório que a reflexão sobre consciência fonológica precisa fazer parte da formação inicial e continuada do(a) professor(a) alfabetizador(a), para que ele(a) tenha condições de realizar intervenções adequadas no processo de aquisição da leitura e da escrita, diagnosticadas por meio da PB – Leitura; e isso foi enfatizado no terceiro encontro, momento em que os(as) professores(as) analisaram as sugestões pedagógicas contidas na fundamentação teórico-metodológica da PB – Leitura.

No segundo encontro, o propósito foi o de identificar as habilidades descritas na Matriz de Referência para Avaliação da PB – Leitura; estabelecer a relação entre matriz de avaliação e matriz curricular; apresentar as relações entre consciência fonológica e alfabetização; refletir sobre as habilidades de leitura que compõem a Matriz de Referência para Avaliação da PB – Leitura; e propor possíveis habilidades que possam ser avaliadas em um TAI.

Para esse encontro, foi proposta a leitura antecipada do documento Brasil (2016a), que identificava os cinco eixos estruturantes das habilidades imprescindíveis para o desenvolvimento da alfabetização e do letramento inicial, quais sejam: 1) apropriação do sistema de escrita; 2) leitura; 3) escrita; 4) compreensão e valorização da cultura escrita; e 5) desenvolvimento da oralidade. O documento esclarecia também que, em função das características da PB, a matriz de referência considerava habilidades apenas dos eixos 1, 2 e 4, sendo que este último não se caracterizava como um eixo separado, mas permeava toda elaboração do teste. Assim, a matriz de referência é apresetada apenas nos eixos 1 e 2. Muitos(as) professores(as) revelaram não saber da organização da matriz em dois grandes eixos, permeados por um terceiro ou que

cada eixo continha a descrição das habilidades condutoras da construção dos itens do teste, apesar de terem aplicado o teste em diferentes momentos de sua prática e de a aplicação ser exigência oficial na rede municipal desde a edição de 2008 da PB.

O documento também ressaltava que o trabalho de desenvolvimento dessas habilidades, durante o processo de ensino e de aprendizagem, não acontecia de maneira sequencial e linear, embora a disposição das habilidades na estrutura dessa matriz seja apresentada linearmente. A compreensão desse aspecto por parte do(a) professor(a) alfabetizador(a) é de fundamental relevância para que não haja confusão entre a matriz curricular e a matriz de avaliação, evitando reducionismos nos processos de ensino e de aprendizagem, ou seja, a matriz de avaliação não pode determinar aquilo que será ensinado ao longo do processo de alfabetização, pois é um recorte do currículo. Essa compreensão também foi fundamental para verificar as limitações do teste diagnóstico da proficiência em leitura em relação aos processos de alfabetização e de letramento inicial, entendidos de forma mais ampla, elucidando que, mesmo com limites, o teste possibilita informação sobre aspectos considerados relevantes desses processos.

Segundo Morais (2012), o Sistema de Escrita Alfabética da Língua Portuguesa, falada no Brasil, apresenta dez propriedades, quais sejam:

> 1. Escreve-se com letras, que não podem ser inventadas, que têm um repertório finito e que são diferentes de números e de outros símbolos.
> 2. As letras têm formatos fixos e pequenas variações produzem mudanças na identidade das mesmas (p, q, b, d), embora uma letra assuma formatos variados (P, p, *P*, *p*).
> 3. A ordem das letras no interior da palavra não pode ser mudada.
> 4. Uma letra pode se repetir no interior de uma palavra e em diferentes palavras, ao mesmo tempo em que distintas palavras compartilham as mesmas letras.
> 5. Nem todas as letras podem ocupar certas posições no interior das palavras e nem todas as letras podem vir juntas de quaisquer outras.
> 6. As letras notam ou substituem a pauta sonora das palavras que pronunciamos e nunca levam em conta as características físicas ou funcionais dos referentes que substituem.
> 7. As letras notam segmentos sonoros menores que as sílabas orais que pronunciamos.
> 8. As letras têm valores sonoros fixos, apesar de muitas terem mais de um valor sonoro e certos sons poderem ser notados com mais de uma letra.
> 9. Além de letras, na escrita de palavras, usam-se, também, algumas marcas (acentos) que podem modificar a tonicidade ou o som das letras ou sílabas onde aparecem.
> 10. As sílabas podem variar quanto às combinações entre consoantes e vogais (CV, CCV, CVV, CVC, V, VC, VCC, CCVCC...), mas a

> estrutura predominante no português é a sílaba CV (consoante – vogal), e todas as sílabas do português contêm, ao menos, uma vogal.

Durante esse segundo encontro de formação, os docentes manifestaram dificuldades em relacionar as habilidades descritas na matriz da PB – Leitura a essas propriedades, bem como identificar sua relevância no currículo da alfabetização. Nas discussões, muitos(as) professores(as) também destacaram a necessidade de avaliar a produção escrita do(a) aluno(a), ao que foi esclarecido que essa importância é reconhecida nos documentos Brasil (2015a, 2015b), porém, o teste de escrita foi suspenso por dificuldades de ordem metodológica. Somente a edição de 2008 contou com o teste de escrita.

No segundo encontro, os(as) professores(as) ainda se referiram à necessidade de restringir os gêneros textuais abordados nos itens aos que eram comumente trabalhados no Ciclo de Alfabetização[42], por exemplo, contos tradicionais no lugar de pequenos artigos de divulgação científica. Eles comentaram que os contos tradicionais eram mais fáceis de serem trabalhados nos anos iniciais do ensino fundamental do que os artigos de divulgação científica. Em relação à escolha dos gêneros textuais que deveriam servir de base para os itens de uma avaliação externa, Morais, Leal e Pessoa (2013) disseram que é necessário discutir e chegar a acordos sobre quais gêneros textuais parecem mais adequados a uma avaliação padronizada, qual deve ser a extensão máxima dos textos usados nos itens, que nível de complexidade seria o mais adequado etc. Dessa maneira, definir quais textos eram mais relevantes para compor uma avaliação externa depende de uma ampla discussão entre os professores(as) alfabetizadores(as) e os(as) elaboradores(as) das referidas avaliações.

Ainda nesse segundo encontro foi discutida a relação entre currículo e avaliação e revelou-se a existência de três desafios, que só poderiam ser enfrentados com o desenvolvimento de processos de formação sobre avaliações padronizadas: a compreensão por parte dos(as) professores(as) sobre os pressupostos que embasam as habilidades descritas na matriz; a relevância desses pressupostos para os processos de alfabetização e de letramento inicial e a necessidade de ampla

[42] O Ciclo de Alfabetização abrange o 1º, 2º e 3º anos do ensino fundamental de 9 anos.

discussão sobre o currículo praticado no final da educação infantil e nos anos iniciais do ensino fundamental.

Quanto à reflexão sobre construção de um TAI da PB – Leitura igualmente realizada nesse segundo encontro, ponderou-se que parte das dificuldades encontradas nos testes em papel e lápis poderiam perfeitamente ser superadas por meio de um dispositivo eletrônico, por exemplo, em relação à avaliação da escrita, poderia ser proposta uma atividade para identificar, em um conjunto de letras, quais compõem determinada palavra ouvida (ditada pelo dispositivo) ou escrever uma palavra ouvida (ditada pelo dispositivo). Por ser um teste que apresenta itens mais ajustados ao desempenho dos(as) respondentes, o TAI poderia ainda evitar o constrangimento de estudantes com maior proficiência leitora em responder a itens que exigiam leituras de menor complexidade. Da mesma maneira, evitaria que os(as) alunos(as) com baixas proficiências ficassem constrangidos por serem submetidos a itens cujas habilidades exigidas são de leitura parcial ou global de um texto.

O terceiro encontro de formação teve por objetivos: identificar os níveis de leitura adotados pela PB – Leitura e estabelecer relações entre a descrição dos níveis de proficiência, em termos dos domínios alcançados, as habilidades da matriz de referência e as sugestões de atividades propostas para a intervenção.

Segundo Brasil (2015a), as respostas dos(as) alunos(as) podem ser interpretadas estabelecendo-se uma relação entre o número ou a média de acertos de um(a) ou mais alunos(as) e sua correspondência com os níveis de desempenho descritos para a PB – Leitura. Dessa forma, o(a) estudante demonstrará ter desenvolvido determinada proficiência quando conseguir responder corretamente a um quantitativo de questões do teste. Para a PB – Leitura são definidos e descritos cinco níveis de alfabetização e letramento inicial. Na análise da descrição de cada um dos cinco níveis, os(as) professores(as) apontaram questionamentos sobre o processo de alocação dos(as) alunos(as) nos níveis, pois alegavam estar analisando os resultados apenas do ponto de vista quantitativo, ou seja, contar os acertos dos(as) alunos(as) e alocá-los, de acordo com a quantidade de acertos, em um determinado nível de desempenho. Foi necessário discutir brevemente sobre como o teste configura uma escala de proficiência, cujos itens traduzem o domínio de tarefas cognitivas por parte de um grupo de respondentes e que os acertos estão relacionados às interpretações mais gerais dos níveis de desempenho. Os(as) professores(as) estavam

incomodados, por exemplo, com a possibilidade de um(a) aluno(a) acertar ao acaso itens difíceis, totalizando acertos que o(a) alocaria em níveis de domínio superior, não correspondendo ao domínio real do(a) aluno(a). A esse respeito, foi mostrado que, mesmo acertando algum item difícil aleatoriamente, isso não aconteceria com muita frequência, fazendo com que o(a) aluno(a) ainda apresentasse um total de acertos correspondente ao domínio de menor proficiência, pois o nível de proficiência prevê um intervalo de acertos. Sendo assim, com poucos acertos aleatórios, ele ainda estaria no intervalo interpretado como sendo de baixa proficiência leitora. Vale destacar que, durante o encontro de uma das escolas participantes da pesquisa, a coordenadora pedagógica relatou que a análise dos resultados da PB – Leitura do ano anterior corroborava essa explicação, visto que ela analisou atentamente o desempenho de sua escola na PB – Leitura e cotejou esses resultados com outras sondagens realizadas pelos(as) professores(as), encontrando intensa equivalência entre os níveis de proficiência dos resultados da PB – Leitura e outros diagnósticos realizados pelos(as) professores(as).

Nesse encontro, também foi proposta uma reflexão sobre as relações entre os níveis de desempenho e as sugestões de intervenção apresentadas em Brasil (2015a). Nesse momento, os(as) professores(as) criticaram algumas dessas sugestões, ou porque, pela experiência, já se valiam delas e se tornariam repetitivas, ou porque perceberam que eram pouco eficazes para a realização das intervenções pedagógicas junto aos(às) alunos(as) das turmas de alfabetização. Como exemplo, para o nível 1 de desempenho, o documento apresenta sugestão para a realização da leitura diária do alfabeto e construção de listas, mas os(as) professores(as). ressaltaram que essas atividades já são feitas exaustivamente no 1º ano e, em compensação, sugerem que outras práticas deveriam ser indicadas, como localizar nomes e letras em gêneros textuais mais simples. Em relação a essas observações, percebeu-se que o foco das sugestões poderia avançar tanto em relação ao tipo de atividade quanto nas reflexões que o(a) professor(a) poderá realizar com os(as) alunos(as) a partir dessas mesmas sugestões. Trata-se de, além de indicar a construção de listas ou leituras diárias, proporcionar reflexões sobre às propriedades do sistema de escrita alfabética, considerando as dificuldades de cada estudante.

Nesse terceiro encontro, revelou-se que algumas equipes gestoras se voltaram para uma análise mais qualitativa dos resultados da PB – Leitura, com vistas a definir

intervenções necessárias para que determinado grupo de estudantes avance, em termos de aprendizagens, dentro de um nível ou para um nível subsequente.

Conforme exposto, nesse processo de formação evidenciaram-se dificuldades de professores(as) e gestores(as) na utilização da interpretação pedagógica da PB – Leitura, corroborando as afirmações de Morais, Leal e Pessoa (2013), ao alegarem que a escala de desempenho adotada na PB – Leitura pouco auxilia os docentes na sua tarefa de ensinar de forma ajustada às necessidades dos(as) alunos(as) e que, infelizmente, estimula apenas um uso que se encerra na classificação dos níveis de proficiência dos(as) alunos(as). De fato, em alguns casos, os(as) professores(as) apenas determinam o nível em que um grupo de alunos(as) está alocado sem, contudo, definir o que será necessário para que esse grupo avance para níveis superiores da escala apresentada, prática que seria fundamental para que a avaliação formativa fosse efetivada. Esse aspecto é indicativo essencial da necessária participação dos alfabetizadores nas diferentes etapas de concepção, aplicação do teste, interpretação e uso dos resultados e, quanto a isso, verificou-se a ausência de uma política de formação que incorpore a discussão sobre medida educacional como suporte para a avaliação formativa.

A formação foi realizada nas 15 unidades por um grupo de formadores que produziu pautas comuns e conduziu discussões de forma articulada. A síntese das discussões que foi apresentada refletiu os apontamentos de campo da autora, referente ao acompanhamento dos encontros em 7 escolas, e os relatos dos formadores envolvidos, em reunião de síntese geral do trabalho realizado.

A síntese apresentada é reveladora dos desafios a serem enfrentados na formação em avaliação, que são: a compreensão por parte dos(as) professores(as) sobre os pressupostos que embasam as habilidades descritas na matriz de referência e sua relevância para o processo de alfabetização, iniciando pela apropriação do disposto nos manuais e documentos da PB – Leitura; a necessidade de ampla discussão sobre o currículo praticado no final da Educação Infantil e nos anos iniciais do Ensino Fundamental e a relação destes com a avaliação; e a problematização das escalas e seu papel de apoio aos processos avaliativos, com o envolvimento de gestores(as) e professores(as) na efetivação de planos de ação, visando ao atendimento às diferentes necessidades dos alfabetizandos.

Além dos desafios revelados na formação, as reflexões e críticas denotaram as preocupações com cada aspecto da medida educacional, envolvido na ferramenta, em especial, que ela seria um meio para apoiar a realização de uma prática avaliativa que continuaria sob a responsabilidade dos(as) professores(as).

4.3 Aplicação do TBC da PB – Leitura

Vários autores apresentam fases/etapas para a construção de um TAI (Cf. BARRADA, 2012; MAGIS; MAHALINGAM, 2015; MOREIRA JUNIOR, 2011; OLEA et al., 1999; OLEA; PONSODA; PRIETO, 1999; OLEA; PONSODA, 2003; PASQUALI, 2013) e revelam, também, a preocupação com a existência de um banco de itens parametrizados conforme a TRI, pois essa modelagem nos permite montar testes individuais adaptados ao nível de realização de cada aluno(a) e garante pontuações comparáveis entre os diferentes testes.

Para Pasquali (2013), a construção desses bancos é laboriosa e demorada, levando cerca de três anos para se concretizar, pois exige a produção de itens, a pré-testagem em amostra representativa, a análise psicométrica para obtenção das características de bons itens e manutenção periódica que visa tanto a ampliação do banco como a verificação da permanência dessas características para os itens do banco. A análise psicométrica de itens por meio da TRI facilita a constituição do acervo de itens, pois

> [...] ela permite estabelecer os parâmetros dos itens [...] independente da amostra de sujeitos utilizada, daí é possível incluir sempre novos itens diretamente comparáveis com os já inclusos no banco [...]. A técnica para esta façanha, entre outras, consiste em aplicar os novos itens juntamente com uma amostra de itens já incluídos no banco a uma amostra razoável de sujeitos e estimar os parâmetros dos novos itens em confronto com os dos itens utilizados no banco de itens. (PASQUALI, 2013, p. 282)

Para o desenvolvimento do Projeto TAI da PB – Leitura, o Inep disponibilizou, além de planilhas com os parâmetros dos itens[43], os seguintes relatórios:

[43] O Inep forneceu planilha com os parâmetros dos itens da edição 2015 da PB – Leitura.

a) Documento técnico B – Material psicométrico para oficina de interpretação de escalas da PB, referente aos itens criados em 2008, 2009 e 2010 (BRASIL, [2011b]);
b) Metodologia de validação e estruturação dos parâmetros psicométricos das edições de 2010 e 2011 da PB (BRASIL, 2011a);
c) Relatório com as Estatísticas e Análise Pedagógica dos Itens. Primeiro Pré-teste de itens do Banco Nacional de Itens (BNI) do Inep 2012 – subgrupo 1(BRASIL, 2012c);
d) Relatório da análise clássica e TRI dos itens. Pré-teste 2013 do Banco Nacional de Itens (BNI) do Inep – subgrupo 1 (BRASIL, 2013a);
e) Relatório da análise clássica e TRI dos itens. Pré-teste 2014 do Banco Nacional de Itens (BNI) do Inep – subgrupo 1 (BRASIL, 2014).

As análises psicométricas encontradas nos relatórios fornecidos pelo Inep são pautadas em duas concepções que fundamentam a análise dos testes: a Teoria Clássica dos Testes (TCT) ou Teoria Tradicional dos Testes (TTT) ou ainda Teoria Clássica das Medidas (TCM) e a denominada Teoria da Resposta ao Item (TRI), na qual o traço latente é o construto principal. Embora cada concepção se apoie em modelos diferentes, ambas são importantes. Segundo Pasquali (2013), a psicometria procura explicar o sentido que têm as respostas dadas pelos sujeitos a uma série de tarefas, tipicamente chamadas de itens e busca parâmetros para essa explicação.

A Psicometria assume os postulados da teoria da medida, que busca expressar por variáveis os comportamentos ou características humanas intangíveis como depressão, inteligência, qualidade de vida, aptidões, entre outros comportamentos que são observados – na verdade, inferidos – indiretamente por respostas aos itens do teste.

Os parâmetros do conjunto de itens da PB – Leitura, encontrados nos relatórios, foram obtidos a partir de amostras representativas da população brasileira. Cabe esclarecer que nem todos os relatórios fornecidos pelo Inep detalharam as análises pela TCT, mas todos descreviam as análises por meio da TRI.

A construção do teste informatizado da PB – Leitura adotou diretrizes indicadas em Andrade, Tavares e Valle (2000), Muñiz(1997), Muñiz e Hambleton (1999), Olea, Ponsoda e Prieto (1999), Pasquali (2007,2013), Lord (1977), entre outros, e as

informações dos Relatórios Técnicos fornecidos em Brasil (2011a, [2011b], 2012c, 2013a e 2014), que explicavam o perfil psicométrico dos testes.

Com base nas diretrizes para que os TAI sejam bons testes, Muñiz e Hambleton (1999) salientaram a importância de se verificar: dimensionalidade, validade, confiabilidade e estimação dos parâmetros dos itens. Essa preocupação coincide com a Etapa 1 indicada por Moreira Junior (2011, p. 153-154), a qual "permite verificar a existência prévia do teste e definir a dimensão, o traço latente e o objetivo do teste".

4.3.1 O objetivo do teste e o traço latente aferido no TBC e no TAI da PB – Leitura

A PB – Leitura é um teste existente na versão impressa desde 2008, concebido e constituído pelo Inep, órgão de âmbito federal, vinculado ao Ministério da Educação (MEC) brasileiro. Quanto à finalidade do teste, a PB surgiu no contexto de ampliação progressiva do Ensino Fundamental para nove anos com o ingresso das crianças de seis anos de idade na escolarização obrigatória, conforme apontam as legislações e documentos indicados no sítio do referido órgão[44]. A ampliação do ensino fundamental para nove anos ocorreu em razão, também, dos resultados alarmantes do Sistema de Avaliação da Educação Básica (Saeb), cujas coletas iniciaram em meados da década de 1990.

Para o Governo Federal, essa ampliação possibilitaria maior tempo para apropriação de conteúdos relativos à alfabetização e ao letramento inicial (BRASIL, 2006), concretizando a indicação do ensino obrigatório de 9 anos, a iniciar-se aos 6 anos de idade. A Lei nº 10.172 (BRASIL, 2001), que aprovou o Plano Nacional da Educação (PNE), transformou essa ampliação em meta da educação nacional. O ensino obrigatório brasileiro também foi modificado (Cf. BRASIL, 2009), definindo que é dever do estado brasileiro garantir a "Educação Básica obrigatória e gratuita dos 4 aos 17 anos de idade".

A PB foi instituída por meio de Portaria Normativa nº 10/07 (BRASIL, 2007), na qual se atribui ao Inep a responsabilidade por estruturar e conceber essa avaliação, fornecendo às secretarias de educação o instrumento e material de instruções para a

---

[44] Portal Inep disponível em:< http://portal.inep.gov.br/provinha-brasil>. Acesso em: 15 jan. 2018.

aplicação. Conforme informa o sítio oficial do Inep, a primeira aplicação impressa atingiu 22 unidades federativas e 3.133 municípios brasileiros.

No período de 2008 a 2011, a adesão não estava diretamente vinculada aos programas de alfabetização, contudo, em 2012, com a instituição do Pacto Nacional para Alfabetização na Idade Certa (Pnaic) (BRASIL, 2012a), a PB se vinculou aos processos de monitoramento da alfabetização, uma vez que sua aplicação passou a vigorar nas incumbências previstas para a federação, os estados, o Distrito Federal e os municípios. Decorrente do Pacto, foi publicada a Portaria nº 387 (BRASIL, 2015c), que estabeleceu normas para a adesão das redes educacionais, dando publicidade à lista de municípios dessa adesão.

Para uma ideia da abrangência da PB após o Pnaic, na lista publicada em 2016 aparecem 5.571 municípios, dos quais 5.474 apresentam público-alvo para a PB e 4.343 indicaram a adesão, totalizando 79,33% dos municípios do Brasil.

Ela foi uma ferramenta de uso disseminado entre professores(as) dos anos iniciais do ensino fundamental e, no início de 2015, foram distribuídos 2.643.187 (dois milhões, seiscentos e quarenta e três mil, cento e oitenta e sete) instrumentos (CATALANI; TATAGIBA, 2015). Essa disseminação não seria possível se a produção dessa versão do teste não dialogasse diretamente com professores(as) alfabetizadores(as) nas escolas, conferindo validade nacional à PB.

A PB possibilitou que os(as) próprios(as) professores(as) aplicassem e verificassem os acertos dos(as) alunos(as), permitindo que eles(as) fossem alocados em níveis da escala de proficiência, que recebeu quatro cortes, definidores de cinco níveis de proficiência. Dessa forma, verificou-se que o teste apresentava características de uma avaliação formativa, pois admitia a classificação dos estudantes em níveis de proficiência, com interpretação pedagógica e sugestão de intervenção.

Em meados de agosto de 2016, os testes não foram mais distribuídos para os municípios, ficando apenas disponíveis em documentos no formato PDF no sítio do Inep e, a partir de 2017, também não foram mais disponibilizados nesse sítio, ainda que a Portaria Normativa nº 10/07 e outras legislações relacionadas não tenham sido revogadas. Uma nota foi publicada (TOKARNIA, 2016) justificando a suspensão do teste por restrições financeiras e “até que sejam publicadas novas Matrizes de Referência para Avaliação da Alfabetização”.

Até o momento de fechamento deste trabalho, a Portaria que instituiu a PB não tinha sido revogada, mas a publicação da Resolução CNE/CP nº 2 (BRASIL, 2017) definiu uma base curricular comum para educação infantil e ensino fundamental no país, com alterações significativas no tempo destinado para alfabetização e letramento inicial, as quais certamente terão implicações para o teste da PB – Leitura.

A PB – Leitura se apoia nos pressupostos de medição de um fenômeno que não pode ser observado diretamente, como acontece com a aprendizagem, a depressão, a qualidade de vida, entre outros. A medida educacional e as medidas psicossociais, resguardados os questionamentos existentes entre os psicólogos, apoiam-se em modelos usados na medição de atributos não extensivos ou não observáveis, definidos como traços latentes. Recorre-se, nesses casos, à observação do comportamento em tarefas, que são denominadas itens.

A portaria de criação da PB já apontava alguns aspectos da dimensão aferida no teste, contudo, são os documentos de orientação para a aplicação e interpretação dos resultados da PB, oferecidos pelo Inep, que especificam o construto ou traço latente na forma de uma matriz de avaliação para cada área de conhecimento. A produção do TAI focalizou a proficiência em leitura em Língua Portuguesa, portanto, apenas a matriz da PB – Leitura será abordada.

A proficiência leitora é tratada como uma competência (traço latente) da área de Língua Portuguesa e as habilidades a serem desenvolvidas para a conquista dessa competência são expressas na Matriz de Referência para Avaliação da Alfabetização e do Letramento Inicial. A matriz de referência, reproduzida no Anexo A, norteia a elaboração de itens e está dividida em eixos, conforme foi anteriormente citado.

Importante destacar que outro documento disponibilizado pelo Inep sobre a PB, o *Guia de elaboração de itens da PB*, também, indica a matriz de referência e estabelece que as habilidades dessa matriz estão fundamentadas

> […] na concepção de que a alfabetização e o letramento são processos a serem desenvolvidos de forma complementar e paralela, entendendo-se a alfabetização como o desenvolvimento da compreensão das regras de funcionamento do sistema de escrita alfabética e o letramento como as possibilidades de usos e funções sociais da linguagem escrita, isto é, o processo de inserção e participação dos sujeitos na cultura escrita. (BRASIL 2012b, p.14)

A próxima etapa da construção do TAI objetivou verificar se o meio eletrônico de administração do teste alteraria o traço latente avaliado, em outras palavras, se os itens poderiam ser utilizados no TAI com suas respectivas calibrações, por considerar que o meio eletrônico não alteraria a escala. Para tanto, foi formulada uma versão eletrônica da PB – Leitura, denominada TBC da PB – Leitura, cujas características são descritas na subseção a seguir.

4.3.2 As características da plataforma e a aplicação do TBC da PB – Leitura

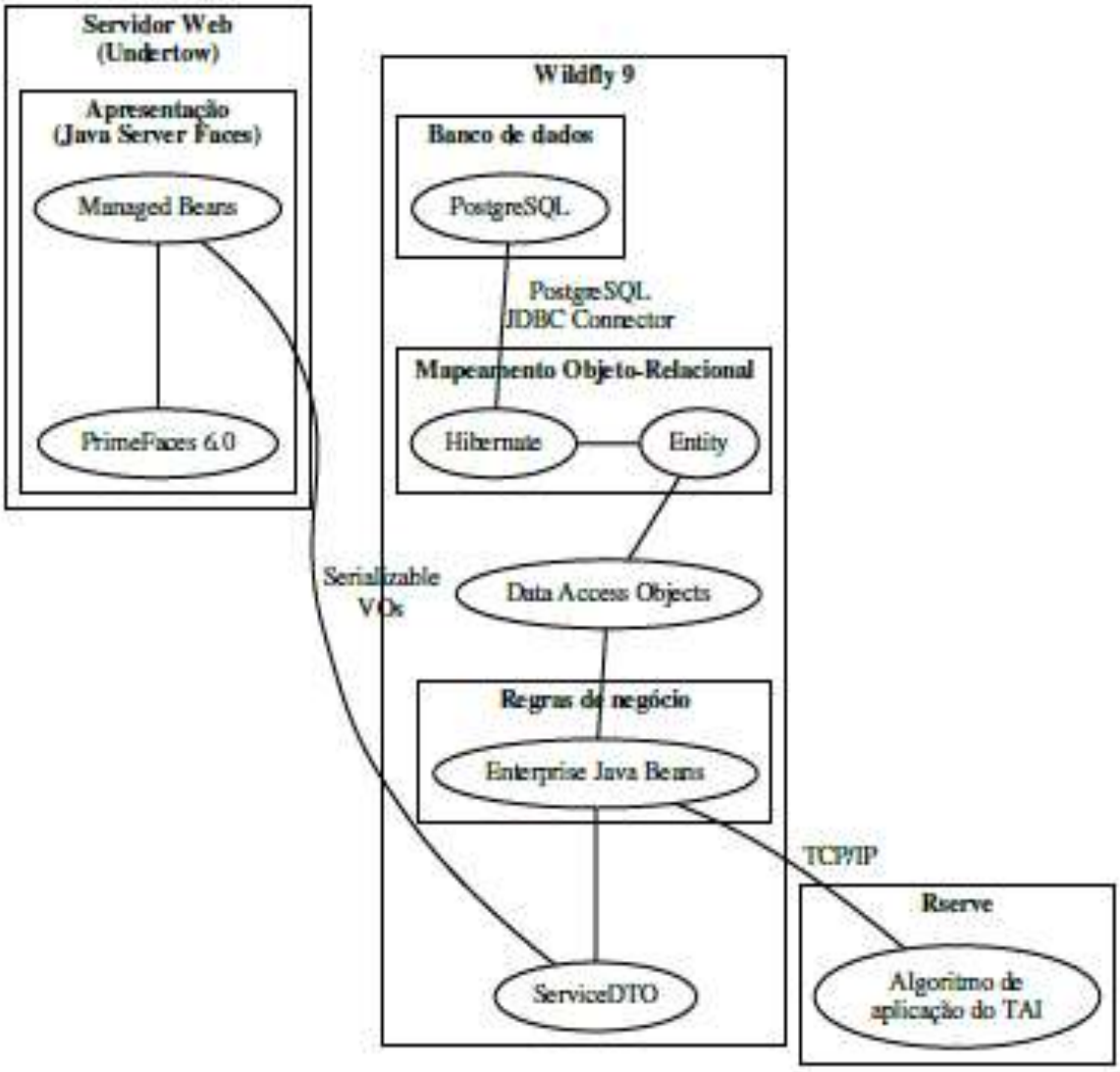
A descrição do *software* para aplicação do TBC da PB – Leitura foi embasada nos relatórios São Paulo (2016a, 2016b). O *software* foi criado em plataforma WEB, garantindo que qualquer dispositivo com navegador de Internet pudesse ser utilizado para aplicação do teste. Para garantia da comparabilidade das aplicações no papel e lápis e no TBC da PB – Leitura, uma das necessidades do sistema era a exibição da totalidade do item na tela do dispositivo, garantindo a mesma apresentação do caderno impresso, seja na posição horizontal seja na posição vertical do *tablet*. O(a) examinando(a) deveria visualizar todos os elementos do item na tela do *tablet*: o número de ordem, o enunciado, todas as alternativas e o botão para avançar para o próximo item.

Para a administração do teste *on-line*, foi proposta uma aplicação que se baseou em um sistema apresentado por Meneghetti e Destro (2012), contendo as funcionalidades básicas como: cadastro de usuários, itens e aplicação de provas aos(às) alunos(as), suporte de apresentação de texto, imagens e áudio, substituindo a narração do aplicador do teste da PB – Leitura. Essa aplicação também tinha a possibilidade de apresentar, imediatamente ao término do teste, o relatório dos resultados ao(à) próprio(a) estudante ou aos(as) professores(as) e gestores(as).

Esse sistema em plataforma WEB foi atualizado e utilizava a linguagem de programação primária Java 1.8 e a plataforma *Java Enterprise Edition* (J2EE) em camadas em conjunto com o servidor de WEB *Undertow* e com servidor de aplicação *Wildfly* 9. Foram utilizados diversos padrões de desenvolvimento e *frameworks*, como os *Enterprise Java Beans* (EJB), mapeamento objeto-relacional com *Hibernate* 5.1, interfaces WEB via *PrimeFaces* 6.0 e a *Build Maven* 3.1. O desenvolvimento da aplicação seguiu o desenvolvimento em camadas, proporcionando independência

entre os *frameworks* utilizados em cada camada. O banco de dados utilizado foi Postgre SQL 5.1. A Figura 5 (SÃO PAULO, 2016a) ilustra a arquitetura do sistema, incluindo o módulo do algoritmo que foi posteriormente agregado.

Figura 5 – Arquitetura do projeto TAI da PB – Leitura e relação entre as diferentes plataformas



Fonte: São Paulo (2016a, p. 7)

A apresentação dos itens na tela do *tablet* necessitou de reuniões entre o assessor responsável e os pesquisadores envolvidos no projeto para análise das telas e indicações quanto às configurações mais adequadas aos(às) alunos(as).

O sistema para o TBC da PB – Leitura necessitou da seguinte infraestrutura nas escolas para ser utilizado: uma conexão internet, um dispositivo móvel ou computador

com conexão e capacidade de reprodução de áudio, um navegador WEB atualizado (*Google Chrome* ou *Opera Safari*) e um fone de ouvido.

Em setembro de 2016, foi realizado um teste de infraestrutura e do aplicativo em uma das escolas, para verificar o acesso simultâneo ao *Wi-Fi* da escola por múltiplos *tablets*, assim como o comportamento dos servidores da SME que disponibilizavam o aplicativo em uma aplicação simultânea. Participaram do teste 12 professores(as) e seis voluntários(as), alguns usando mais de um *tablet*. Os(as) professores(as) foram orientados sobre como usar o aplicativo e realizaram uma ou mais vezes o TBC da PB – Leitura, construído com os itens do teste 2 da edição de 2016 da PB – Leitura. Exceto por problemas relacionados à atualização dos *tablets*, que estava sendo feita gradualmente nas escolas, todos os testes foram realizados com sucesso.

Os(as) professores(as) e gestores(as) aprovaram o tamanho das fontes tipográficas e das imagens e a distribuição do conteúdo na tela do *tablet*. Puderam testar a síntese de voz, relatando que eram compreensíveis. Eles deram uma estimativa para o tempo da prova eletrônica: em duas aulas, ou seja, 1 hora e meia, e criticaram o limite de vezes – duas – que o(a) aluno(a) poderia ouvir a síntese de voz que fazia a leitura do enunciado de cada item, defendendo que deveria ser maior. Quanto a essa limitação, foi explicado que era fundamental para manter as mesmas características da aplicação do teste da PB – Leitura em papel e lápis.

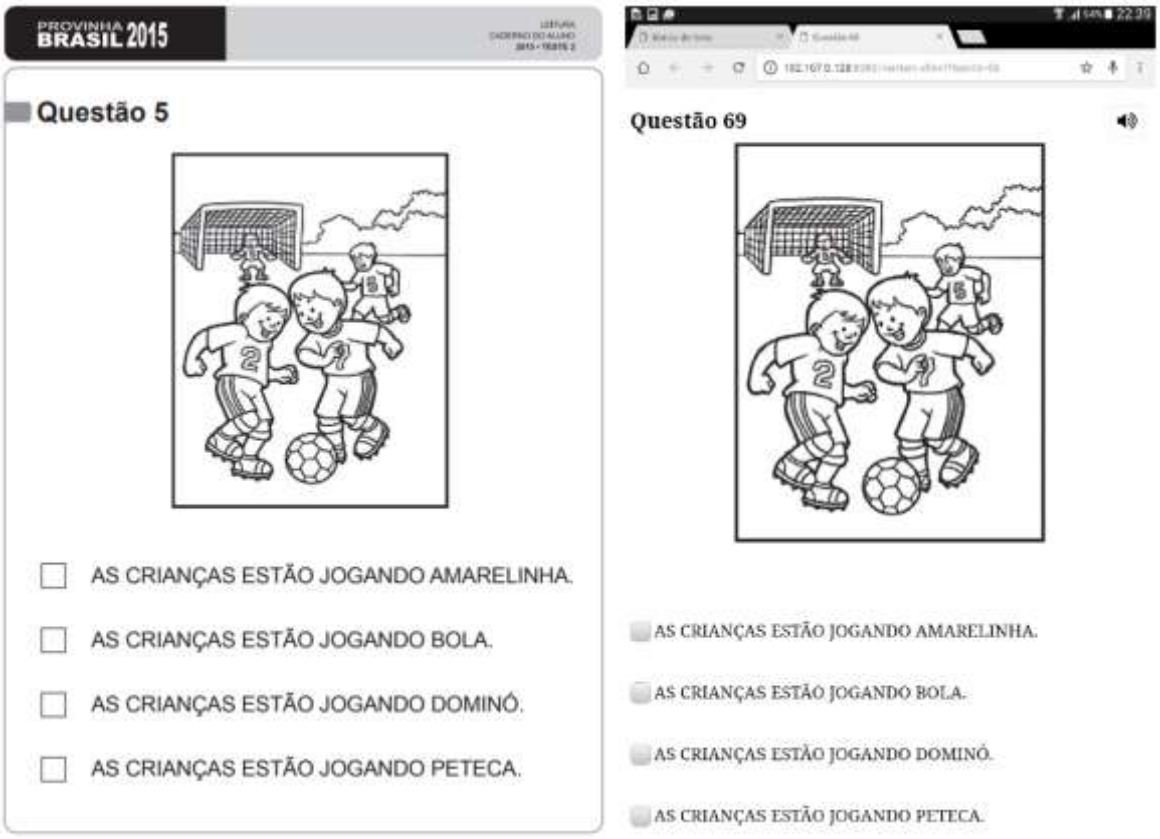
Os relatórios São Paulo (2016a, 2016b) apontam que, durante a aplicação do TBC da PB – Leitura, o uso dos servidores alcançou a média de 5% de processamento e uso constante de 3 GB de memória RAM. A conexão de rede de 8 MB/s da escola teve pico de uso de 1,5 MB/s e média de 500 kB/s.

As aplicações de TBC da PB – Leitura foram realizadas com a utilização de *tablets* da Samsung Galaxy Tab 10.1, existentes nas escolas e na SME/SP, e o navegador Google Chrome, versão 49.

Além da equivalência no número de vezes que os(as) alunos(as) poderiam ouvir o item no TBC da PB – Leitura, era necessário garantir a equivalência do aspecto visual do item na prova impressa e na tela do *tablet*, incluindo a correspondência da visualização do item, na disposição vertical ou horizontal da tela do *tablet* (Figura 6). A transposição dos itens para a tela do *tablet* foi feita utilizando o material disponibilizado pelo Inep em seu portal *on-line*. O material era composto de arquivos

em formato PDF – Caderno do aluno e Caderno do aplicador –, sendo que, do Caderno do aluno, foram retiradas as imagens e os textos que seriam mostrados na tela do *tablet,* e do Caderno do aplicador, os textos que seriam "ditados" eletronicamente para os(as) estudantes.

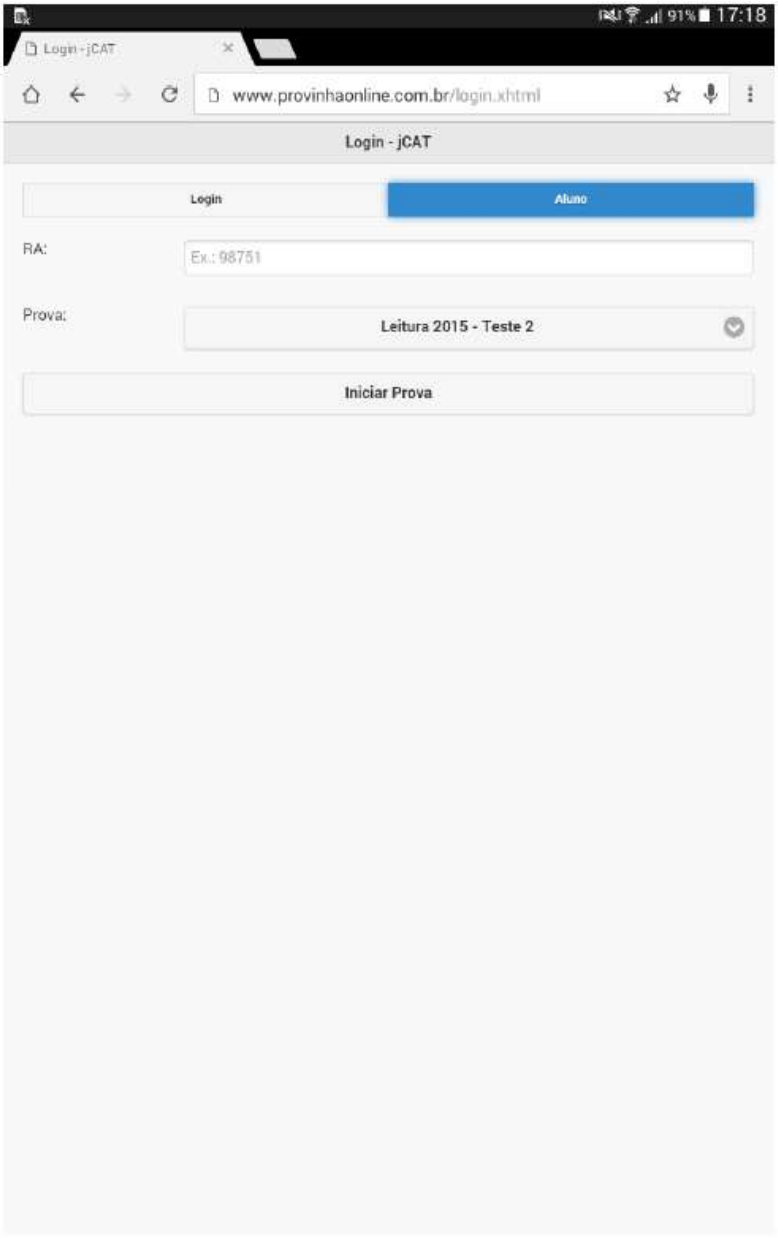
Figura 6 – Item 5 da PB – Leitura, nas formas impressa e digital (*tablet)*.



Fontes: Brasil (2016b) e São Paulo (2016a).

Os *tablets* eram entregues para os(as) alunos(as) já ligados e conectados à aplicação do teste eletrônico da PB – Leitura na aba de *login* do "Aluno" (Figura 7). Essa etapa inicial foi orientada por um aplicador(a)/estagiário(a), que fornecia também o número do RA de cada estudante para que fosse digitado no campo de *login*.

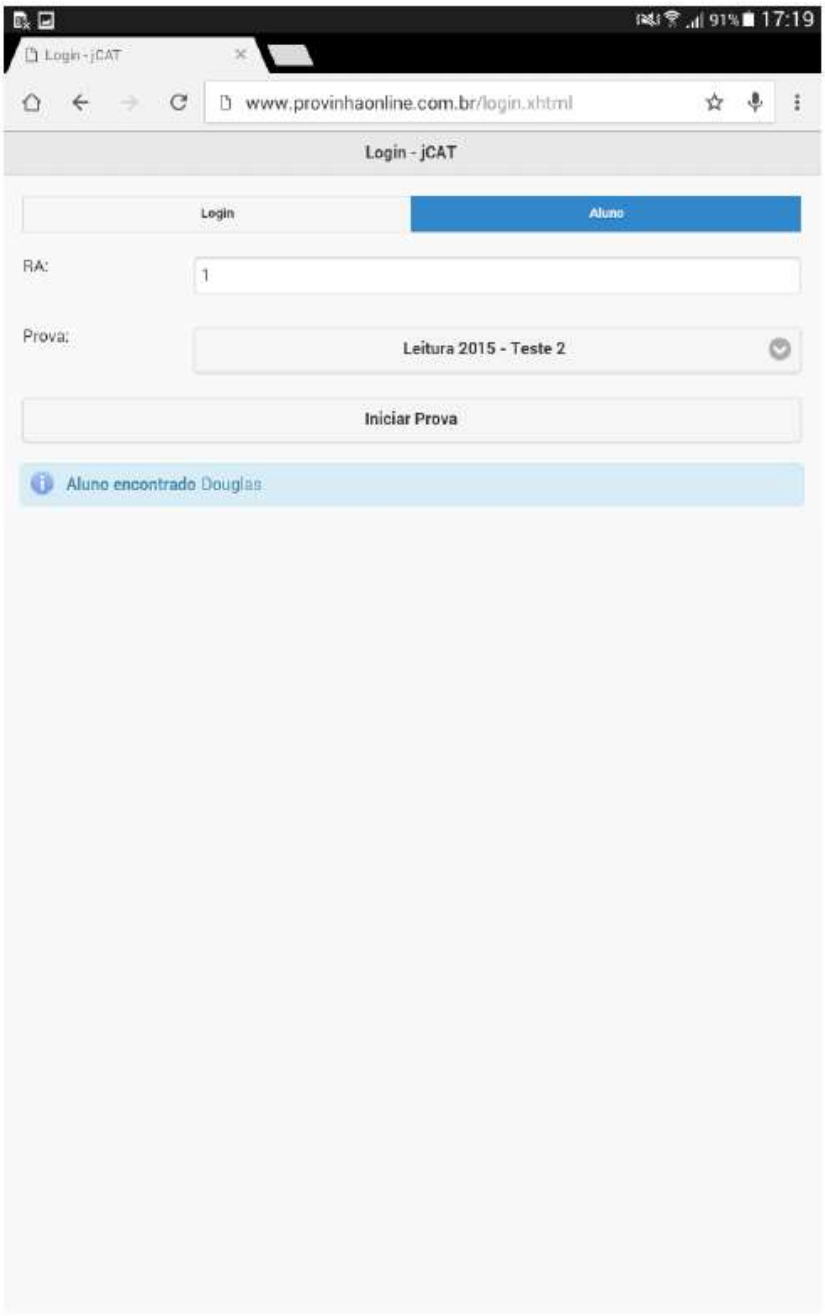
Figura 7 – Tela de *login* no TBC e TAI da PB – Leitura



Fonte: São Paulo (2016a).

Caso o(a) aluno(a) estivesse inserido(a) no sistema, uma confirmação na área inferior do formulário era exibida (Figura 8).

Figura 8 – Mensagem de existência do(a) aluno(a) no sistema TBC e TAI da PB – Leitura



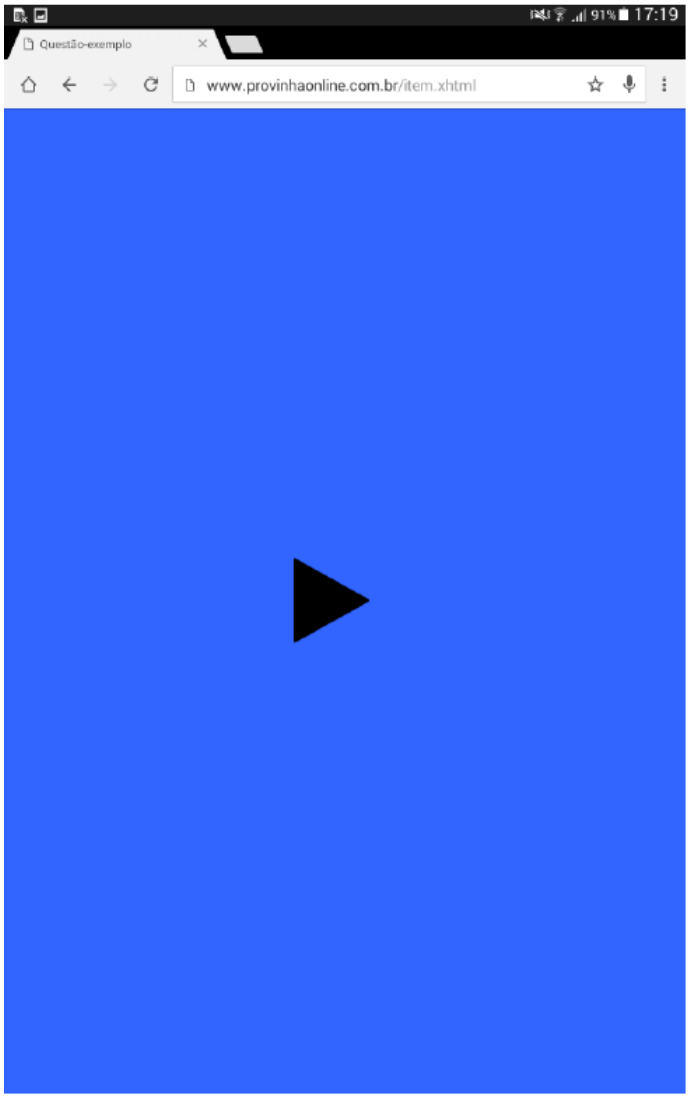
Fonte: São Paulo (2016a).

O aplicador do TBC da PB – Leitura solicitava ao(à) aluno(a) que conferisse se o volume do *tablet* estava habilitado, preferencialmente em seu máximo. O uso de fones de ouvido, principalmente na aplicação coletiva, era necessário para garantir que cada aluno(a) ouvisse apenas o áudio de sua prova. A configuração do volume dos dispositivos e dos fones de ouvido foi uma dificuldade adicional e exigiu o auxílio de

estagiários(as) no momento da aplicação. Outro aspecto que dificultou as aplicações se referiu ao limite de conexões simultâneas dos *tablets* à rede sem fio das escolas, resultando em aplicações para no máximo 16 crianças por sessão. Havia uma heterogeneidade na quantidade e disposição dos pontos de acesso à rede sem fio, bem como na configuração que limitava a quantidade de conexões em cada ponto de acesso, significando outro obstáculo, que requereu remarcação de datas de aplicações de prova e a distribuição dos(as) alunos(as) de uma turma em mais de uma sala.

Em seguida, o(a) aluno(a) selecionava o botão “Iniciar Prova” e era confrontado com uma tela contendo um único botão (Figura 9).

Figura 9 – Tela inicial de início da prova no TBC e no TAI da PB – Leitura



Fonte: São Paulo (2016a).

Tocando no botão, iniciava-se a questão-exemplo (Figura 10). Nela, os principais conceitos da interface de aplicação da prova eram apresentados ao(à) aluno(a) por meio da síntese de voz.

Figura 10 – Tela com questão exemplo do TBC e do TAI da PB – Leitura

Fonte: São Paulo (2016a).

Nessa tela, o(a) aluno(a) recebia orientação guiada quanto aos elementos disponíveis na tela, sendo que cada explicação era acompanhada de uma seta azul pulsante, mostrando o botão de áudio, o qual oferecia a leitura – locução automatizada – do enunciado, as alternativas e o botão para avançar para o próximo item.

O enunciado dos itens era lido automaticamente para o(a) aluno(a), assim que um item era carregado na tela. Ao tocar no botão de áudio  , visível do lado superior direito da tela do *tablet*, conforme mostrado nas Figuras 10 e 11A, o(a) aluno(a) podia ouvir o enunciado uma segunda vez, emulando a aplicação da PB – Leitura convencional, na qual o(a) professor(a)/aplicador(a) lê para o(a) aluno(a). Após a segunda vez, o botão de áudio desaparecia.

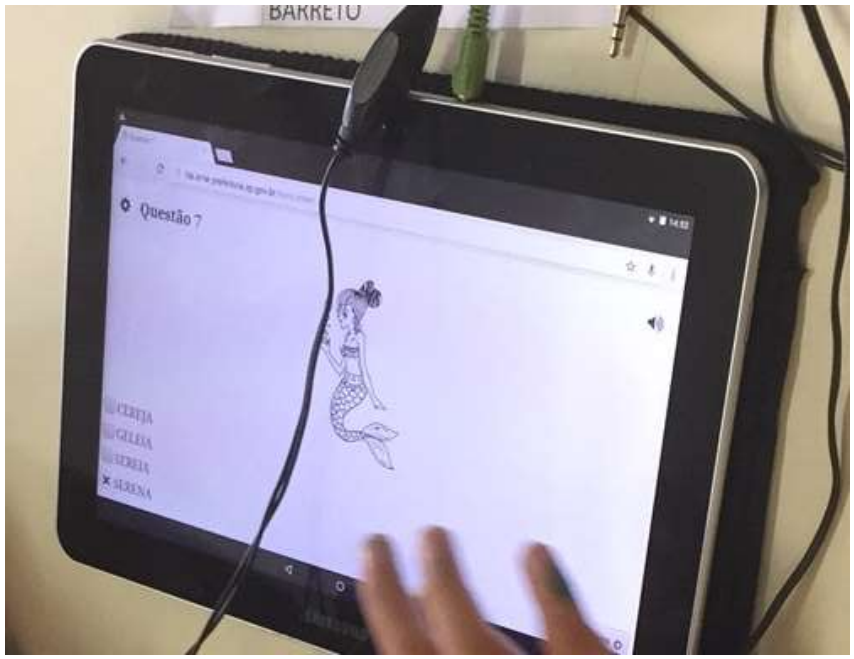
Esse botão de áudio dava acesso a uma locução automatizada dos textos que na prova impressa eram lidos pelo(a) professor(a)/aplicador(a). Para essa locução foi usada a Interface de Programação de Aplicativos[45] (API) de síntese de voz do Google.

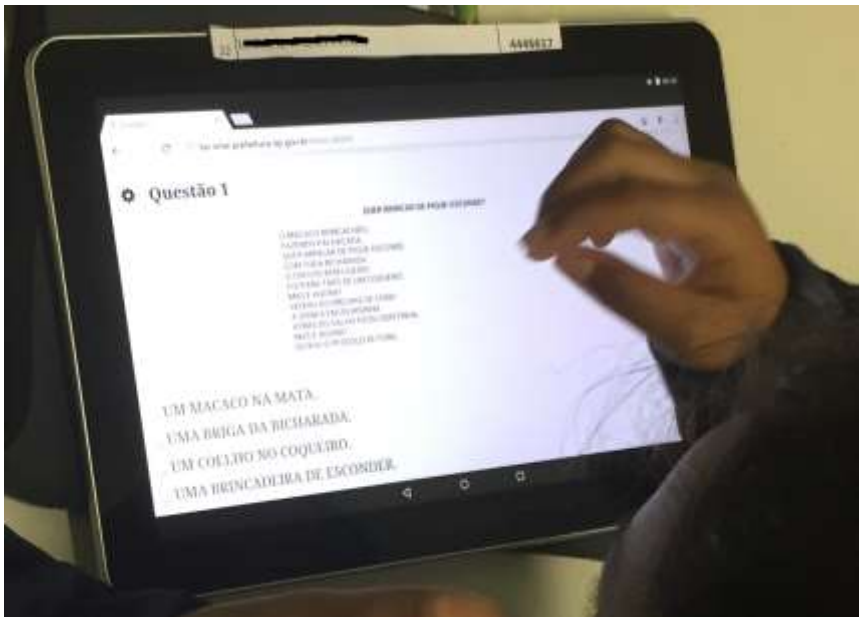
Quanto à leitura do enunciado dos itens, é preciso esclarecer que os itens se dividiam, por apresentarem duas formas diferentes de locução informatizada, aspecto que representou desafio a ser transposto na construção da versão TBC da PB – Leitura. Havia itens que necessitavam de uma única locução e itens que necessitavam de duas locuções. Para os itens que requeriam uma única locução automatizada (Figura 11A), ela era realizada no momento que o texto-base e as alternativas do item eram carregados na tela do *tablet*, conforme explicado anteriormente. Para os itens que requeriam uma dupla locução automatizada (Figura 11B), referindo-se aos itens que possuíam um pedido anterior ao enunciado (pergunta), por exemplo, pediam ao(à) aluno(a) para ler um texto ou uma tirinha, antes de apresentar a pergunta com a tarefa a ser realizada sobre o texto ou a tirinha. Esse pedido anterior era apresentado logo que a primeira parte do item era carregada na tela (não eram carregados as alternativas e o botão para o próximo item). Assim que essa primeira parte era carregada, o botão de áudio iniciava uma animação, indicando que deveria ser tocado. Somente quando tocado, a pergunta era lida para o(a) aluno(a) e as alternativas apresentadas, evitando que o(a) aluno(a) pudesse responder ao item sem ouvir a pergunta. O(a) aluno(a) poderia selecionar a repetição dessa locução ou selecionar a alternativa desejada. Somente depois de selecionada a resposta, o botão de “próximo item” era habilitado.

---

[45] Em inglês, *Application Programming Interface*.

Figura 11 – Apresentação do item com locução única e dupla no TBC da PB – Leitura

A. 

B. 

Fonte: A autora. Foto da aplicação do TBC da PB – Leitura.

.

O dispositivo previa o encerramento automático do teste após transcorridas 2 horas, caso o teste não fosse encerrado pelo estudante por motivos adversos, após ter respondido aos 20 itens.

Na aplicação do TBC, os(as) alunos(as) não poderiam retornar aos itens já respondidos ou avançar para o item seguinte sem assinalar uma resposta. Essa decisão foi tomada para equiparar a aplicação do TBC da PB – Leitura à dinâmica de aplicação da PB – Leitura em papel e lápis, na qual todos(as) os(as) alunos(as) são conduzidos pelo(a) professor(a), que realiza a leitura necessária a cada tipo de item, determinando um tempo médio para as respostas e definindo o avanço de todos os(as) alunos(as) para o próximo item.

Os TBC da PB – Leitura foram aplicados nas 15 escolas participantes do projeto. Em todos os casos, metade dos(as) alunos(as) de cada turma fez a prova eletrônica. Não obstante o sistema TBC da PB – Leitura realizasse a apresentação da interface automaticamente aos(às) examinandos(as) no início da prova, alguns alunos(as) esperavam receber instruções de um aplicador, ou seja, alguns alunos(as) aguardavam por uma confirmação do adulto para prosseguir no teste. Assim, os aplicadores tinham que ficar atentos às telas dos *tablets*, garantindo que todos os(as) alunos(as) estivessem prosseguindo na resolução dos itens, independentemente de o sistema ter sido planejado para que o(a) examinando(a) chegue ao fim do teste sem intervenção de terceiros.

A Tabela 5 mostra a quantidade de provas eletrônicas aplicadas por escola, assim como o horário de início da primeira prova eletrônica e o horário de término da última prova eletrônica. Foram aplicadas 524 provas eletrônicas para os(as) alunos(as); destas, somente 1 não foi validada.

Tabela 5 – Horário inicial e final (em minutos) e número de aplicações do TBC da PB – Leitura, por escola

Escola	Número de aplicações	Início do teste	Fim do teste
Angola	37	14:32	15:04
África do Sul	29	16:49	17:52
Moçambique	32	10:32	11:06
Libéria	35	09:35	11:53
Argélia	25	14:13	16:48
Costa do Marfim	38	14:38	17:39
Líbia	19	16:36	17:00
Cabo Verde	35	14:23	14:43
República do Congo	47	17:08	10:08
Camarões	47	14:20	15:25
Etiópia	26	09:39	10:28
Benim	25	14:01	14:59
Ruanda	40	10:43	16:13
Egito	46	16:53	17:33
Marrocos	43	10:38	12:41
Total	524	-	-

Fonte: São Paulo (2016b), com substituição dos nomes das escolas.

Na Tabela 6, são apresentados os tempos médios para prova e por item nos TBC da PB – Leitura, avaliados por escola, revelando uma média de tempo de aproximadamente 15 minutos para a prova e de 43 segundos para o item. Ressalta-se que se deve considerar que o tempo total de duração do TBC da PB – Leitura para uma turma foi um pouco maior, devido à necessidade de se esperar que todos os(as) alunos(as) de uma turma terminassem de usar os *tablets* para que fosse chamada a turma seguinte. Em algumas escolas, por conta de problemas técnicos da conexão, havia mais de uma sessão de aplicação para uma turma, exigindo o dobro de tempo para as orientações iniciais, dado que os primeiros 5 a 10 minutos eram destinados às instruções básicas para inserção do código do(a) aluno(a) e ajuste do volume dos fones de ouvido. De modo geral, o tempo para aplicação do TBC da PB – Leitura, por turma, foi de 20 a 30 minutos, enquanto o tempo para a aplicação da prova na versão impressa variou de 25 a 60 minutos, com média de 38 minutos.

Tabela 6 – Tempo médio (em minutos) do TBC da PB – Leitura, por escola, por prova e por item

Escola	Tempo médio de prova	Tempo médio por item
Angola	00:17:05	00:00:48
África do Sul	00:15:13	00:00:43
Moçambique	00:18:00	00:00:51
Libéria	00:13:27	00:00:39
Argélia	00:10:54	00:00:31
Costa do Marfim	00:20:50	00:00:59
Líbia	00:13:34	00:00:38
Cabo Verde	00:12:01	00:00:34
Camarões	00:14:01	00:00:40
Etiópia	00:15:54	00:00:45
Benim	00:15:10	00:00:43
República do Congo	00:13:14	00:00:37
Ruanda	00:16:34	00:00:50
Egito	00:12:28	00:00:35
Marrocos	00:16:11	00:00:46
Média	00:14:58	00:00:43

Fonte: São Paulo (2016b), com substituição dos nomes das escolas.

A Tabela 7 sintetiza as informações de tempo para responder aos 20 itens nos dois meios de aplicação, sendo que no TBC foi considerado apenas o tempo para resolução dos itens e não o de orientação. Assim, diferente da Tabela 6, a média foi calculada com base no tempo de cada TBC da PB – Leitura. Os dados de tempo de aplicação para a prova impressa constam do Anexo E. Com base na Tabela 7, nota-se que houve uma diminuição de aproximadamente 23 minutos do teste convencional para o TBC da PB – Leitura, representando uma diminuição de 60,5%.

Tabela 7 – Duração do teste (em minutos), por tipo de administração

Tipo de administração	Tempo por teste (em minutos)		
	Médio	Mínimo	Máximo
Impressa (papel e lápis)	38	25	60
Eletrônica (TBC)	15	2	56

Fonte: São Paulo (2016b).

A Tabela 8 revela a quantidade de estudantes que realizaram o TBC da PB – Leitura, por nível de proficiência. A distribuição dos(as) alunos(as) nos níveis seguiu a regra utilizada para o teste em papel e lápis, representada na Tabela 9.

Tabela 8 – Distribuição dos(as) alunos(as) no TBC da PB – Leitura, por nível da escala e por escola

Escola	Nível 1	Nível 2	Nível 3	Nível 4	Nível 5
Angola	0	8	18	9	2
África do Sul	0	9	16	4	0
Moçambique	0	9	19	4	0
Libéria	0	8	24	3	2
Argélia	1	6	13	7	1
Costa do Marfim	2	6	21	9	0
Líbia	0	3	12	2	2
Cabo Verde	1	5	21	7	1
República do Congo	0	16	23	7	1
Camarões	1	5	29	12	0
Etiópia	1	3	13	7	2
Benim	1	5	13	5	1
Ruanda	2	9	17	10	4
Egito	1	12	31	1	1
Marrocos	1	17	19	5	3
Total	11	121	289	92	20
%	2,06	22,7	54,22	17,26	3,75

Fonte: São Paulo (2016b), com substituição dos nomes das escolas.

Pode-se perceber que há uma distribuição que se assemelha à distribuição normal dos(as) alunos(as) entre os níveis de proficiência, sendo que o nível 3 contém a média dos níveis de proficiência e há quantidade equivalente de alunos(as) nos níveis 2 e 4 e nos níveis 1 e 5.

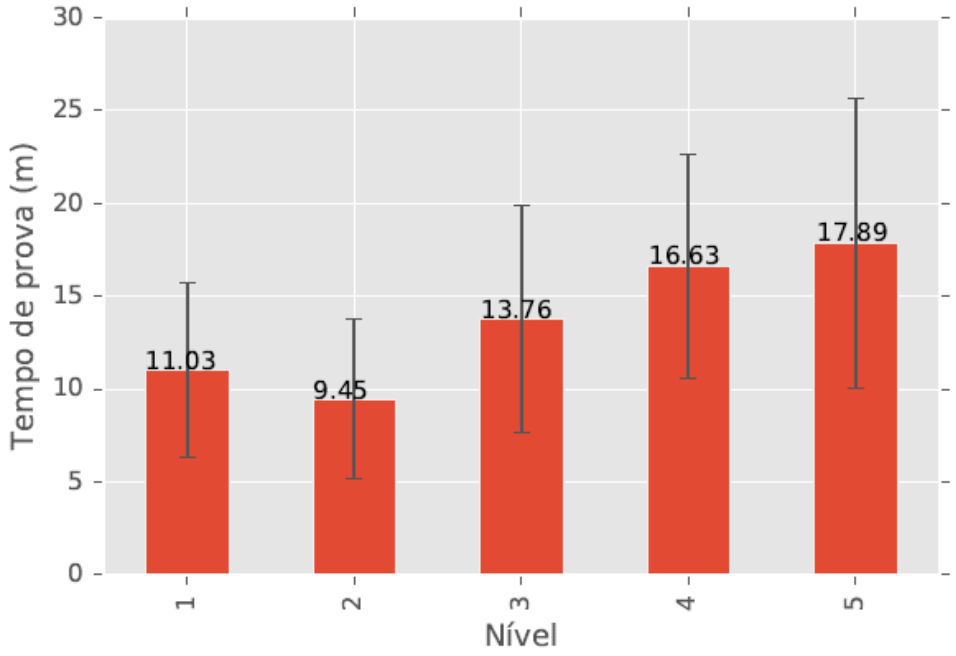
Tabela 9 – Relação entre acertos e níveis de desempenho no Teste 2 da PB – Leitura, edição 2016

Nível de desempenho	Total de acertos
Nível 1	até 2
Nível 2	de 3 a 7
Nível 3	de 8 a 13
Nível 4	de 14 a 15
Nível 5	de 16 a 20

Fonte: Brasil (2016c, p. 24).

A Figura 12 exibe a média de tempo, em minutos, para a conclusão da prova para alunos(as) dos diferentes níveis da escala. É possível observar que alunos(as) de menor proficiência (níveis 1 e 2) levaram tempo menor para concluir a prova, enquanto os(as) alunos(as) de maior proficiência levaram, em média, maior tempo.

Figura 12 – Média e desvio padrão para o tempo de conclusão do TBC da PB – Leitura (em minutos), por nível de proficiência



Fonte: São Paulo (2016b).

Uma explicação para essa diferença se deve ao fato de os(as) alunos(as) com maior proficiência, alocados(as) nos níveis 4 e 5, já apresentarem domínio de leitura e desempenharem a leitura dos textos dos itens mais difíceis, levando maior tempo nessa tarefa. Outros aspectos sobre o tempo de prova serão analisados em comparação com o TAI da PB – Leitura, em seção seguinte.

Adicionalmente, ressaltam-se outras vantagens que o TBC da PB – Leitura apresenta em relação à aplicação em papel e lápis: a) o transporte das respostas para uma folha de respostas, aspecto que adquire especial importância para respondentes de determinadas faixas etárias, por exemplo, crianças por serem inexperientes na realização de testes; b) a padronização da leitura do item quando indicada e a possibilidade de cada aluno(a) seguir com o teste ao seu ritmo; e c) digitalização das respostas para constituição do banco de dados para análise estatística, pois a interação dos(as) respondentes com o dispositivo eletrônico permite que as respostas passem para o banco de dados de forma automatizada, conferindo rapidez na obtenção dos resultados.

Embora os TBC possam permitir a incorporação de ferramentas tecnológicas na elaboração dos itens, diversificando e ampliando as tarefas ou problemas propostos aos(às) respondentes, os itens utilizados no TBC da PB – Leitura não incorporaram essa possibilidade no protótipo construído para a investigação, porque o ponto de partida foram os itens da prova impressa.

Nas observações de Scalise e Gifford (2006) e Sireci e Zenisky (2006, 2016), itens construídos com base em tecnologias podem lançar mão das ferramentas tecnológicas de duas diferentes maneiras: a) nos modos de apresentar os contextos e/ou objetos auxiliares/suportes na reflexão proposta pelo item e que mobilizam uma resposta do(a) respondente, podendo integrar multimídias e agregar movimento e som às figuras, gráficos, textos e ilustrações já utilizados nos itens de testes de papel e lápis; e b) nas operações cognitivas solicitadas, suplantando as possibilidades de expressar escolhas, descrições, identificações, comparações, relacionamentos, análise e avaliações em torno dos fatos, fenômenos, ou linguagens, objetos da aferição pretendida.

De modo adicional, os(as) alunos(as) submetidos(as) ao TBC da PB – Leitura se mostraram muito motivados(as) com o teste no *tablet*, revelando enorme familiaridade com o dispositivo. A diretora de uma das unidades, que participou ativamente da aplicação informatizada, manifestou sua surpresa com a facilidade com que os(as) alunos(as) daquela faixa etária entendiam as orientações e utilizavam o dispositivo. Esse aspecto foi mais enfaticamente observado na aplicação do TAI da PB – Leitura, por envolver os(as) alunos(as) das turmas dos 2º anos na totalidade e parte dos(as) alunos(as) do 1º ano.

Os resultados do TBC da PB – Leitura foram essenciais para o protótipo do TAI da PB – Leitura, aspecto que será abordado na subseção seguinte.

4.3.3 O funcionamento diferencial dos itens segundo o modo de administração da PB – Leitura

O funcionamento diferencial do item (DIF)[46] é utilizado pelos psicometristas para verificar se itens funcionam de maneira igual ou diferente para subgrupos da

[46] Da expressão em inglês *Differential Item Functioning*.

população e a existência de diferenças corresponderia a medir um objeto, usando instrumentos enviesados (MUÑIZ, 1997). Desse modo, o DIF é analisado em testes convencionais, com o propósito de atribuir maior validade e confiabilidade à medida realizada. Costumeiramente, são avaliados efeitos referentes ao nível socioeconômico ou quanto ao gênero (sexo) dos(as) alunos(as) e nos relatórios Brasil (2012c, 2013, 2014) foi apontado que os itens que apresentaram DIF significativo no pré-teste da PB – Leitura foram excluídos do BI.

Entretanto, o DIF também pode ser usado para verificar a equivalência dos parâmetros dos itens do teste impresso para o teste informatizado, conforme argumentam Olea et al. (1999), se a única alteração no teste informatizado foi na forma de apresentação, do papel para o computador, não deveriam ser alteradas as propriedades psicométricas, pois o conteúdo do teste continua o mesmo.

Avaliar a existência ou não de comportamentos diferenciais nos itens significaria poder atribuir os mesmos parâmetros dos itens da versão impressa do teste para a versão informatizada e posteriormente para a versão adaptativa. Essa possibilidade permitiria considerar para o TAI a mesma escala de proficiência da PB – Leitura e dotar de sentido pedagógico o resultado do teste informatizado.

Com essa finalidade, foi planejada a verificação de funcionamento diferencial dos itens (DIF), comparando a versão convencional (em papel e lápis) e um teste com características semelhantes, mas administrado no meio digital, no caso um *tablet* (GONZÁLEZ BETANZOS, 2011). No entanto, a comparação entre as aplicações em papel e meio eletrônico tinha como objetivo central a comparação entre os parâmetros de dificuldade dos itens – o parâmetro b – para cada meio de aplicação, considerando que o TAI estava sendo delineado precipuamente para utilizar a escala de proficiência do teste convencional e garantir o que se considerava essencial: apoiar as decisões decorrentes da prática avaliativa.

Para tanto, os(as) alunos(as) das turmas do 2º ano das 15 escolas foram divididos(as) aleatoriamente e, para uma das metades (523 estudantes), foi aplicado o teste 2 em papel e lápis da edição de 2016 da PB – Leitura e, para a outra metade, foi aplicado o mesmo teste, na versão eletrônica (TBC). Com a aplicação do TBC da PB – Leitura, buscava-se observar a existência de DIF na análise do funcionamento dos itens para o subgrupo que fez a prova em papel e lápis e para o que a realizou no *tablet*. Se todas as condições da prova fossem mantidas e a única mudança ocorresse no meio

de administração do teste – digital versus impresso –, a diferença de funcionamento seria atribuída a isso e, em consequência, os parâmetros dos itens pré-testados no meio impresso não poderiam ser considerados para o novo meio de administração do teste. A necessidade de considerar os mesmos parâmetros se justificava na possibilidade de utilizar a mesma escala de resultados.

É importante lembrar que um TBC, embora apresentado em plataforma informatizada, conforme suas especificações, pode continuar com as características do teste em papel e lápis, quais sejam, apresentar uma sequência fixa e idêntica de 20 itens para todos os(as) respondentes, garantindo ao máximo que as mesmas condições estejam presentes nos dois testes.

Em suas discussões González Betanzos (2011), Magis et al. (2010), Muñiz (1997), Muñiz e Hambleton (1999) e Sisto (2006) têm relativizado a afirmação da presença de viés na verificação de DIF significativo, ao considerar que os índices estatísticos de DIF serviriam para identificar os itens com funcionamento diferencial em distintos grupos e apenas com base em julgamento realizado por especialistas no construto aferido pelos itens é que se poderia determinar quais deles estariam enviesados.

Para os TAI, a análise de itens com DIF permite documentar a existência da equivalência entre as pontuações provenientes das aplicações convencionais e informatizadas. Em geral os procedimentos para detectar o DIF se preocupam em distinguir os itens que definem o traço a ser medido, busca-se identificar quais itens são enviesados e por qual subgrupo, considerando um grupo de referência, por exemplo, a prova informatizada, e um grupo focal, por exemplo, a prova impressa.

A Tabela 10 mostra os resultados de DIF, calculados pelo assessor do NTA da SME/SP, para os 20 itens da PB – Leitura. Dos diferentes métodos de cálculo do DIF, apresenta-se a detecção por meio do método de Mantel-Haenszel (MH), descrito em Magis et al. (2010), no qual foram identificadas diferenças entre moderadas e grandes para os itens 1, 2, 6, 7, 8, 11, 12 e 15. Esse método é o mais popular e destina-se a comparar a resposta dada ao item (correta e incorreta) em cada um dos grupos (focal ou referência) de forma associada à pontuação total do teste, ou seja, para qualquer item testado, são cruzados em uma tabela de contingência 2 x 2 o tipo de resposta (correta ou incorreta) e o número de respondentes de cada grupo. Um item é, portanto, classificado com DIF se o valor da estatística MH for maior que um valor crítico com base na distribuição qui-quadrado. A dimensão do efeito DIF é dada por uma

estatística alternativa, o Δ_{MH} (MAGIS et al., 2010). Para valores de $|\Delta_{MH}| \leq 1$, o efeito é considerado insignificante; se $1 \leq |\Delta_{MH}| \leq 1,5$, o efeito é moderado e se $|\Delta_{MH}| \geq 1,5$, o efeito é grande, sendo que essas interpretações dos efeitos estão baseadas na Escala delta (ETS) de Holland e Thayer (MAGIS et al., 2010). Se os valores são positivos, o item favoreceu o grupo focal (prova impressa) e se negativo favoreceu o grupo de referência (prova informatizada).

Tabela 10 – Resultado de DIF para o TBC da PB – Leitura, pelo método de Mantel-Haenszel (MH)

Acerto ao Item	MH	Δ_{MH}
1	24,139	2,992
2	150,846	5,69
3	1,216	-0,941
4	0,004	0,189
5	0,029	-0,192
6	15,283	-2,279
7	22,164	-2,432
8	14,761	1,786
9	2,086	0,546
10	0,345	-0,331
11	29,901	-1,803
12	11,585	-1,776
13	11,326	-1,385
14	2,872	0,676
15	19,489	1,683
16	5,387	-0,803
17	0,378	-0,243
18	1,16	0,433
19	6,663	-0,982
20	4,422	-0,793

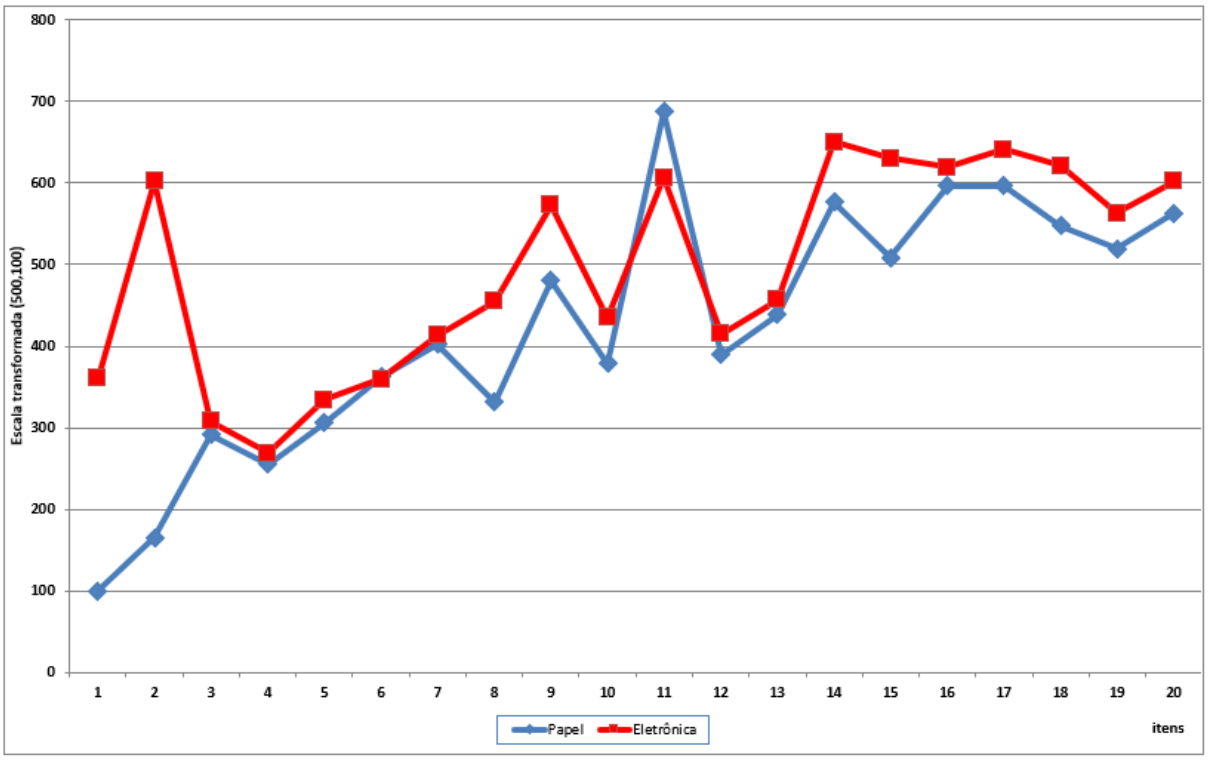
Fonte: São Paulo (2016c).

Embora essa técnica de análise apontasse vários itens com DIF, ela não diz nada a cerca da natureza ou causa do funcionamento diferenciado (MUÑIZ, 1997, p. 160). A análise dos especialistas poderá identificar as causas que originam esse funcionamento diferencial.

Além disso, o problema central era quanto aos parâmetros de dificuldade “b”, pois, na aplicação do TAI, se constituía no elemento mais relevante para o funcionamento do algoritmo, responsável por determinar tanto a sequência de itens a serem administrados quanto o encerramento do teste. Assim, era decisivo avaliar se os parâmetros estimados na versão em papel poderiam ser empregados no TAI e, para isso, efetuou-se uma comparação entre os parâmetros de dificuldade estimados (com

o *software* XCALIBRE) para a aplicação em papel e para o TBC da PB – Leitura, expostos na Figura 13.

Figura 13 – Parâmetros de dificuldade “b” para as aplicações impressa e eletrônica da PB – Leitura



Fonte: Gepave, com base nos microdados fornecidos pela SME/SP.

Foram analisados os parâmetros “b” obtidos para os itens nas aplicações, porque efetivamente era o que interessava para a decisão de usar ou não os parâmetros estimados pelo Inep na aplicação do TAI da PB – Leitura.

É possível observar na Figura 13 que, dos 20 itens, somente os itens 1, 2, 8 e 15 apresentaram diferenças significativas nos valores de “b” aferidos.

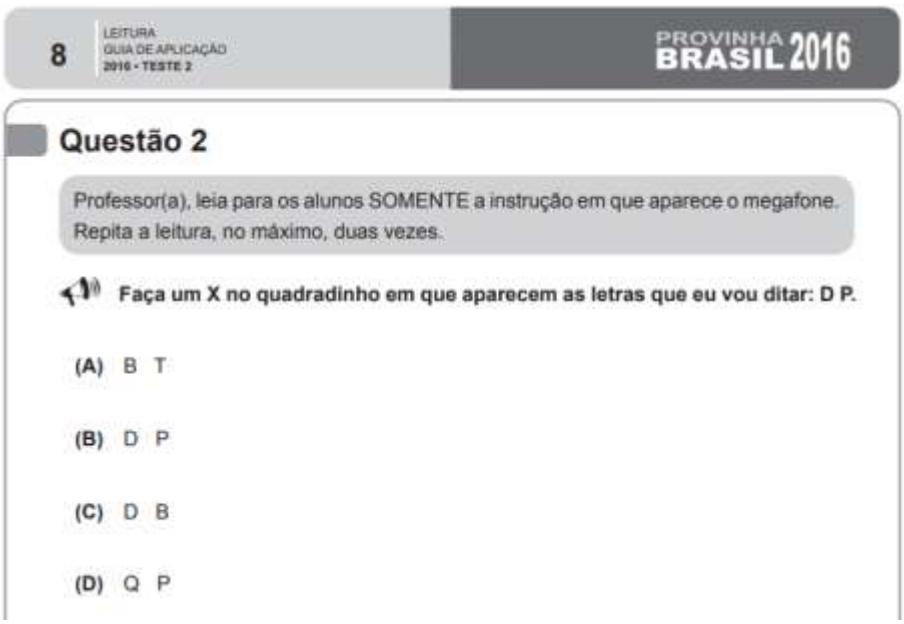
Também foi levado em conta que os(as) estudantes estavam mais familiarizados(as) com os ditados de seus(suas) professores(as) e no TBC tiveram o primeiro contato com a locução eletrônica, podendo acioná-la apenas duas vezes – aspecto que teria paralelo com a orientação dada para a aplicação do teste impresso –, entretanto, conforme consta nos relatórios apresentados no Anexo D, em algumas aplicações da versão em papel foi feita a leitura uma terceira vez. Nas observações da aplicação do teste em papel e lápis, apresentadas no Anexo D, ainda surgiram situações em que

os(as) alunos(as) reproduziam respostas de colegas, aspecto coibido na aplicação do TBC, devido às caraterísticas peculiares ao meio eletrônico, ora porque ele possibilitava avanço na sequência de itens de forma individualizada, ora porque o(a) aluno(a) tinha que se concentrar para escutar o som do seu fone de ouvido.

Os itens com DIF ou com diferenças significativas nos valores de “b” aferidos foram submetidos à análise criteriosa dos especialistas em elaboração de itens para alfabetização e letramento inicial, que contou com representantes do NTA e do Gepave. O objetivo da análise era localizar os aspectos que justificam o comportamento diferencial pelo modo de administração do teste. Considerando os itens apontados, apenas para o item 2 os especialistas identificam o modo de administração do teste como causa para o DIF.

Com fulcro nas ponderações sobre a aplicação do teste e na análise dos especialistas, configurou-se o julgamento de que o modo de administração do teste teria afetado somente o item 2, ilustrado na Figura 14, para o qual a locução da versão eletrônica (leitura automatizada) teria prejudicado a compreensão da escuta das letras “D” e “P”, sendo que a leitura com possibilidade de visualizar o(a) professor(a)/aplicador(a) e articular a escuta com a leitura labial teria favorecido o acerto para o caso de o item envolver soletração de letras.

Figura 14 – Item 2 da PB – Leitura, teste 2, edição 2016



Fonte: Brasil (2016a).

Desse modo, concluiu-se que os parâmetros dos itens do teste convencional poderiam ser utilizados na administração do TAI da PB – Leitura e, para evitar definitivamente o problema observado no item 2, foi dada a possibilidade de o(a) respondente repetir maior número de vezes a locução automática no TAI da PB – Leitura e foi observado que o novo BI também não envolveria item com essa característica: o ditado de letras.

4.4 O algoritmo do TAI da PB – Leitura

Após a decisão de considerar os parâmetros dos itens e a respectiva escala de proficiência da prova em papel e lápis no TAI da PB – Leitura, encaminhou-se o processo de elaboração do seu algoritmo.

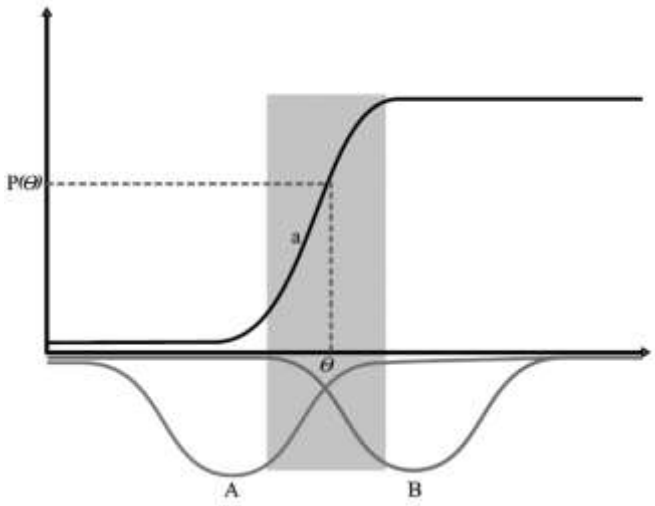
O teste adaptativo oferece testes diferenciados e ajustados aos domínios do(a) respondente. Isso não é uma novidade no campo dos testes, mas a aplicação informatizada e os modelos da TRI tornaram o procedimento muito mais simples, eficiente e rápido. Os modelos da TRI são fundamentais no que concerne ao teste adaptativo, pois apresentam uma característica que possibilita, mesmo com o(a) estudante respondendo a testes diferentes, aferir as proficiências na mesma escala.

O algoritmo de um TAI é o responsável por garantir que cada respondente receba testes diferenciados na quantidade e na complexidade dos itens. Isso é possível porque a seleção dos itens procura aproximar a complexidade do item ao conhecimento do(a) respondente. Mais do que proporcionar desafios possíveis ao(à) respondente, esse procedimento confere maior fidedignidade às estimativas da proficiência. Isso é possível somente com os avanços trazidos pelos modelos da TRI para os processos de medida de um traço latente.

Na TCT, os parâmetros dos itens dependem da população que responde ao teste, sendo que em cada população os itens assumem parâmetros diferentes. Ao contrário, na TRI os parâmetros dos itens possuem a característica da invariância. Em outras palavras, a invariância dos parâmetros dos itens é uma característica central na TRI, ausente na TCT (Cf. COUTO; PRIMI, 2011; OLEA; PONSODA, 2003).

Couto e Primi (2011, p. 10) destacam essa propriedade e ilustram a independência da probabilidade de acertar em relação à amostra de respondentes com a Figura 15.

Figura 15 – Independência dos parâmetros dos itens em relação às amostras A e B da população



Fonte: Couto e Primi (2011).

Na Figura 15, observa-se que a curva característica do item (CCI), existente em função de seus parâmetros, pode ser obtida tanto na amostra A como na amostra B de respondentes, ainda que as distribuições sejam diferentes. Da mesma forma, o valor da proficiência, dada por θ, pode ser obtido utilizando os parâmetros desse item e o θ corresponde à mesma probabilidade de acerto em ambas as amostras.

Em decorrência da invariância dos parâmetros dos itens, esses autores apontaram ser possível utilizá-los para estimar proficiências em novas populações. Os TAI vão depender fundamentalmente da invariância dos parâmetros dos itens, dada pela TRI, na estimação da proficiência de uma pessoa. O TAI da PB – Leitura foi configurado para ter todos os itens com parâmetros, ou seja, calibrados.

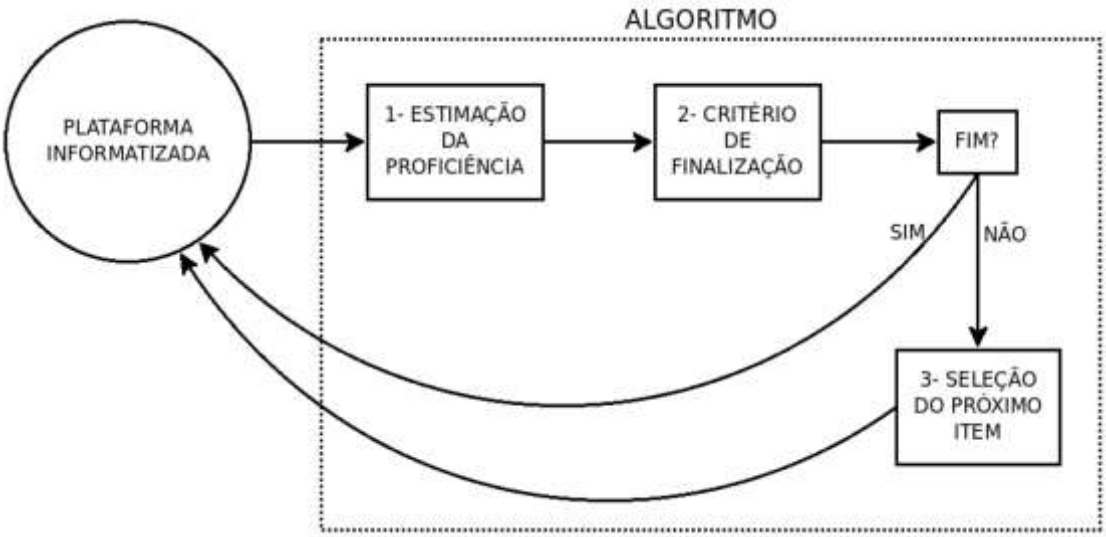
Além de os(as) respondentes receberem itens com níveis de dificuldades sob encomenda para seu nível de proficiência, a quantidade de itens no TAI também pode ser reduzida significativamente em relação aos testes convencionais. Os itens são escolhidos sequencialmente com o objetivo de maximizar o ganho de informação durante uma sessão de testes adaptativos e para que isso seja possível é necessário que as respostas aos conjuntos diferentes de itens permitam estimar proficiências comparáveis, propriedade possível para as técnicas de estimação de teoria de

resposta a itens (TRI). Com os modelos da TRI, o foco da análise passou a ser o item e com as funções de informação tornou-se claro a possibilidade de escolher itens para maximizar a precisão da estimativa da proficiência em qualquer nível da escala.

A descrição do algoritmo do TAI da PB – Leitura foi embasada nos relatórios São Paulo (2016c, 2016d), que assinalaram o acréscimo à plataforma do TBC da PB – Leitura de um algoritmo computacional, para implementar o TAI da PB – Leitura. Esse algoritmo gerenciava a seleção e entrega de itens, a estimação da proficiência e o encerramento do teste.

A Figura 16 ilustra os componentes do TAI PB – Leitura e a inter-relação entre a plataforma construída para o TBC da PB– Leitura e o algoritmo do TAI.

Figura 16 – Esquema geral do algoritmo do TAI da PB – Leitura e seus componentes



Fonte: São Paulo (2016c).

O algoritmo foi programado para usar as respostas dadas aos itens respondidos, a fim de estimar a proficiência e escolher o próximo item com um nível de dificuldade cada vez mais próximo da aferição da proficiência obtida para o(a) respondente.

O formato concreto do algoritmo constituiu-se de um pacote R, normalmente integrado a um ambiente (Linux, Windows ou Mac) que tenha instalado o R na versão 3.0 ou superior. Foram utilizadas algumas funções provenientes dos pacotes catR (MAGIS; RAÎCHE, 2012) e irtoys (PARTCHEV, 2016), devidamente adaptadas aos objetivos do projeto TAI da PB – Leitura.

O algoritmo foi desenvolvido com fundamento na Teoria da Resposta ao Item (TRI), conforme indicam Baker (2001), Muñiz (1999) e Olea e Ponsoda (2002), entre outros e proporcionou uma dinâmica adaptativa à plataforma usada no TBC da PB – Leitura.

De modo específico, o algoritmo foi elaborado para ter três etapas, sendo:

a) estimação de proficiência:
   - recebe como *input* da plataforma as respostas de cada estudante – configurando um padrão de acertos –, e a distribuição *a priori* da proficiência, além dos parâmetros do BI;
   - estima a proficiência e o erro padrão, utilizando um método bayesiano (método estimação *a posteriori* – EAP) e a distribuição *a priori*.

b) critério de finalização do teste:
   - verifica se pelo menos uma das três regras foram alcançadas; se sim, envia um *output* para a plataforma terminar o teste, se não, segue para a terceira etapa. As regras são:
     - o teste alcançou os limites mínimo (7 itens) e máximo de itens (20 itens);
     - a proficiência estimada apresenta erro padrão menor do que o limite máximo definido (35 pontos);
     - verifica a presença de um dos pontos de corte da escala no intervalo de confiança[47] (com nível de 85%) da proficiência estimada.

c) seleção do próximo item do teste:
   - busca o item mais informativo do banco de itens (BI), levando em conta a proficiência estimada na etapa (a); e
   - retorna o item selecionado como *output* para a plataforma.

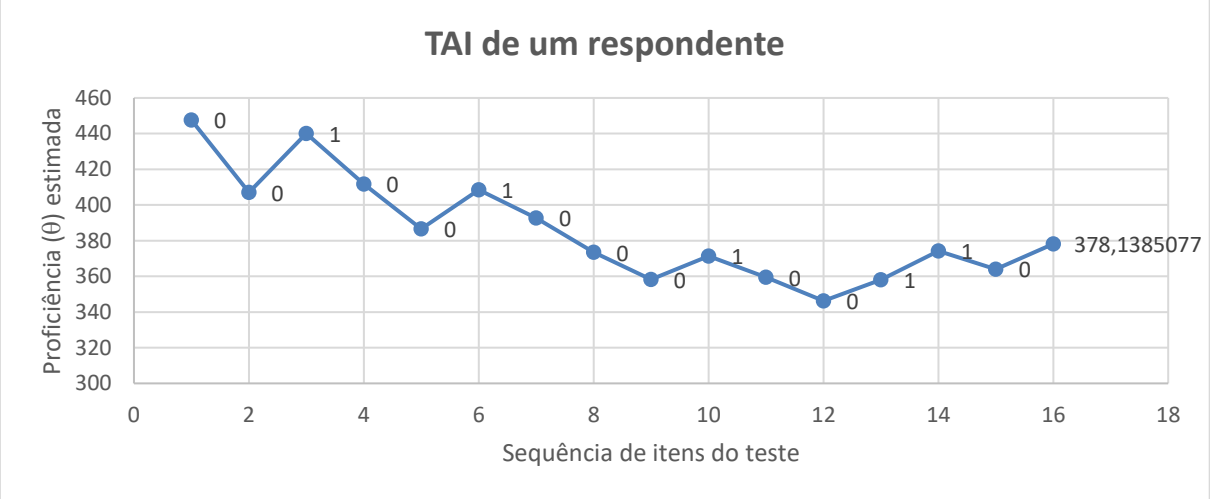
Para dar início à primeira etapa, um item de dificuldade mediana era administrado, ou seja, aquele que se localizava no intervalo central da escala de proficiência da PB –

[47] No desenvolvimento do algoritmo do TAI foi utilizado o intervalo de confiança, quando deveria ter sido utilizado o intervalo de confiança bayesiano (Cf. EHLERS,2007; SCACABAROZI; DINIZ, 2010).

Leitura, cujo parâmetro de dificuldade era próximo de 500[48]. Esse processo de responder o item, estimar a proficiência e verificar o critério de encerramento ocorria continuamente até que este último fosse alcançado.

Segue exemplo de um(a) aluno(a) que respondeu a 16 itens e acertou 6. A proficiência, ao final, foi estimada no valor de 378,1385077, conforme mostra a Figura 17. O erro de medida desse(a) respondente será mostrado na subseção 4.4.3. Os valores 0 e 1 constituem os erros e acertos para os itens apresentados (vetor resposta).

Figura 17 – Itens e respostas de um(a) respondente submetido ao TAI da PB – Leitura



Fonte: A autora, a partir dos dados da aplicação do TAI da PB – Leitura fornecidos pela SME/SP 2016.

A estimação da proficiência foi realizada com base na distribuição esperada a posteriori – (EAP) com 21 pontos de quadratura. O critério para seleção de itens foi inicialmente planejado para incluir duas regras: 1) a máxima informação de Fisher (MFI); e 2) a seleção equilibrada de itens entre os descritores da matriz, ou seja, procura-se manter sempre uma proporção semelhante de itens de cada descritor para garantir que o teste adaptativo represente a matriz desejada. Contudo, a segunda regra não foi contemplada no algoritmo, devido ao restrito BI.

Para determinação do fim de teste, o algoritmo utilizou um critério que combinava três regras:

[48] Referenciada à escala transformada da PB – Leitura, de média 500 e desvio 100.

a) o número de itens do teste (mínimo de 7 e máximo de 20 itens);
b) o limite permitido de incerteza (erro de medida menor que 35 pontos);
c) a classificação da proficiência estimada em um dos níveis da escala da PB – Leitura.

A regra em (c) constituiu uma modificação do critério de finalização do TAI da PB – Leitura, incorporando um critério que normalmente é utilizado em avaliações com finalidade de certificação e que buscam classificar sujeitos em duas ou mais categorias. Essa regra, identificada por regra de classificação, busca verificar se o intervalo de confiança da proficiência aferida está contido no nível da escala. Esse nível, no caso da PB – Leitura, diferencia os domínios de leitura dos(as) alunos(as) e, por conseguinte, apoia o processo de avaliação e de intervenção pedagógica.

As quatro primeiras subseções seguintes trazem os aspectos teórico-metodológicos fundamentais para a constituição do TAI da PB – Leitura, como as especificações do BI, da escala de proficiência do teste impresso e das etapas do algoritmo. A última subseção detalha a aplicação do TAI da PB – Leitura e apresenta análise dos resultados empíricos, que de modo geral são calcados nos relatórios São Paulo (2016c, 2016d) e nos microdados fornecidos pela SME/SP.

4.4.1 Os parâmetros dos itens da PB – Leitura utilizados no TAI da PB – Leitura

Para o TAI da PB – Leitura foi necessária a substituição do BI existente na plataforma do TBC. O banco que continha 20 itens do teste 2 da edição de 2016 foi substituído pelos 40 itens provenientes da edição de 2015 da PB – Leitura, mais precisamente 39 itens, pois um deles apresentou um problema detectado pelos especialistas em elaboração de itens para alfabetização. A substituição do BI do TBC PB – Leitura foi necessária, também porque as mesmas crianças que responderam ao TBC seriam submetidas agora ao TAI da PB – Leitura. Foram usados os itens da edição de 2015 (Tabela 11), porque os parâmetros tinham sido fornecidos pelo Inep.

Tabela 11 – Parâmetros[49] do BI do TAI da PB – Leitura

nº do item	COD.INEP	Parâmetro A'	Parâmetro B'	nº do item	COD.INEP	Parâmetro A'	Parâmetro B'
65	61880	0,0147	341,897	87	36428	0,0147	345,299
66	61571	0,0147	330,770	88	36270	0,0147	335,184
67	62366	0,0147	399,288	89	36143	0,0147	329,652
68	62672	0,0147	374,977	90	62378	0,0147	352,965
69	62584	0,0147	363,277	91	36116	0,0147	361,400
70	62223	0,0147	407,503	92	36221	0,0147	394,571
71	36205	0,0147	448,273	93	62463	0,0147	372,670
72	61678	0,0147	430,487	94	36600	0,0147	379,297
73	62361	0,0147	491,551	95	36697	0,0147	404,369
74	63015	0,0147	517,205	96	36436	0,0147	424,471
75	62924	0,0147	500,091	97	63042	0,0147	457,621
76	36359	0,0147	511,347	98	36388	0,0147	457,510
77	62376	0,0147	470,233	99	62936	0,0147	507,724
78	62107	0,0147	518,711	100	36421	0,0147	489,242
79	61992	0,0147	455,284	101	36190	0,0147	457,481
81	36319	0,0147	605,200	102	14663	0,0147	542,667
82	62404	0,0147	620,113	103	63018	0,0147	575,700
83	36346	0,0147	634,746	104	62994	0,0147	580,758
84	36441	0,0147	571,556	105	14566	0,0147	599,589
86	36533	0,0147	309,542	-	-	-	-

Fonte: São Paulo (2016c).

Os parâmetros dos itens da PB – Leitura foram fornecidos pelo Inep à SME/SP (SÃO PAULO, 2016c) para a realização do estudo. Com base em Brasil [2011b], verificou-se também que, embora na Tabela 11 sejam apresentados dois parâmetros para os itens, o modelo da TRI adotado para a PB foi o logístico unidimensional de 1 parâmetro (ML1P), visto que apenas o parâmetro B é variável. A justificativa para esse aspecto foi a operacionalização do *software*.

Conforme esclarecem Andrade, Tavares e Valle (2000), a TRI propõe modelos para representar a relação entre a probabilidade de um indivíduo dar uma certa resposta a um item e seus traços latentes, ou as características não diretamente observáveis, como a proficiência em leitura, aferida na PB – Leitura.

A proficiência em leitura, traço latente aferido indiretamente por meio do comportamento em tarefas (itens), requer um modelo matemático que permita

[49] Parâmetros na escala transformada, conforme exposto na subseção 4.4.2.

relacionar: traço latente (habilidade) e comportamento (características dos itens). Os modelos da TRI (MUÑIZ, 1999; PASQUALI; PRIMI, 2003) superam as limitações da TCT ao relacionarem a probabilidade de acerto no item e o nível do traço latente, exigindo suposições básicas, como o ajuste dos dados do teste às curvas características dos itens especificadas por um modelo da TRI e a de independência local.

Uma das vantagens desses modelos, sugerem Andrade, Tavares e Valle (2000), é permitir a comparabilidade das proficiências estimadas: para populações submetidas a testes que tenham itens comuns; para indivíduos da mesma população, ainda que tenham sido submetidos a testes totalmente diferentes; e para populações diferentes submetidas a testes totalmente diferentes. Esse potencial de equiparação, característica central na TRI, consiste em ter como elemento principal o item e não o teste na totalidade.

Os itens da PB – Leitura, conforme Brasil (2012c, 2013a, 2014), passaram por pré-testes e as análises indicaram a garantia das suposições básicas supracitadas. Os itens foram tratados como dicotômicos, significando que, embora existam quatro alternativas de respostas (A, B, C e D), considerou-se a resposta apenas como certa (1) ou errada (0).

Os modelos da TRI abordam a probabilidade de um sujeito acertar/aceitar o item em função da proficiência (habilidade) e dos parâmetros que expressam determinadas propriedade dos itens. Para o ML1P, a propriedade é o parâmetro de dificuldade “b”. A fórmula desse modelo indicada em Brasil [2011b, p. 6] é dada por:

Equação 1 – Função da TRI, utilizada na PB – Leitura

$$P(X_{ij} = 1 \mid \theta_j) = \frac{1}{1 + e^{-a_i(\theta_j - b_i)}}$$

Em que:

X_{ij} e uma variável dicotômica que assume valores unitários, quando o respondente *j* responde corretamente o item *i*, ou 0 se não responde corretamente ao item *i;*

θ_j é a proficiência (traço latente) do *j*-ésimo respondente;

$P(X_{ij} = 1 \mid \theta_j)$ é a probabilidade de um(a) respondente *j* com proficiência θ_j acertar/aceitar o item *i;*

a_i é o parâmetro de discriminação do item *i* (no modelo de 1 parâmetro, o valor é comum a todos os itens);

b_i é o parâmetro de dificuldade do item *i,* medido na mesma escala da proficiência;

e é a base do logaritmo natural, um número irracional com número infinito de casas decimais, que será considerado *e* = 2,718;

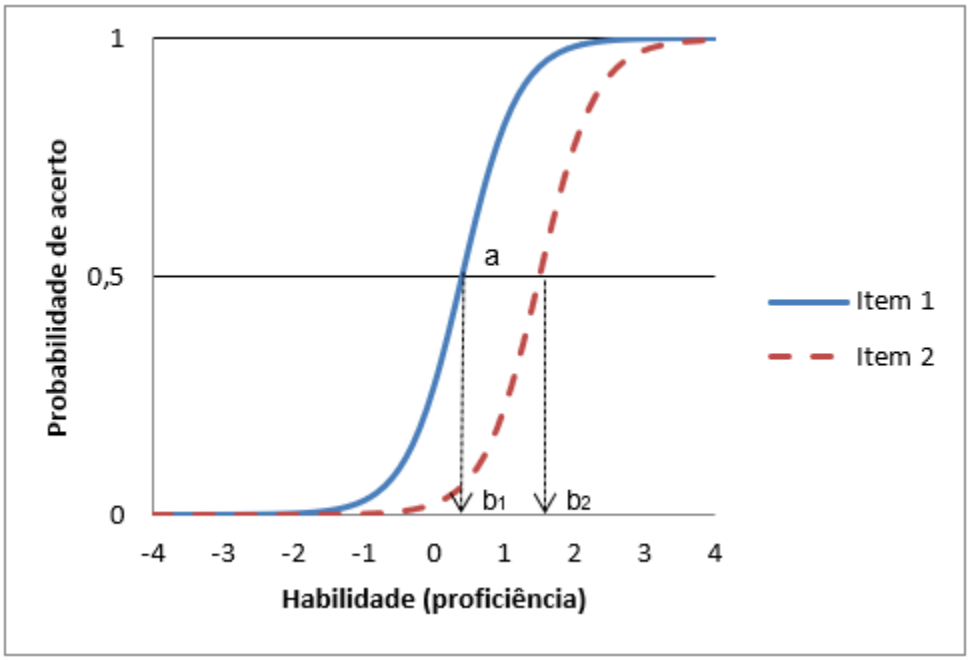
Os modelos da TRI ainda apresentam a vantagem de fornecer os parâmetros dos itens e da proficiência dos(as) respondentes na mesma escala, diferente da TCT. Essa escala é arbitrária e, por padrão, apresenta média 0 e desvio padrão 1. É esperado que os valores das proficiências θ e dos parâmetros b estejam entre – 4 e 4, embora eles possam assumir valores entre $-\infty$ e $+\infty$.

Os parâmetros de dificuldade b dos itens, também são considerados parâmetros da proficiência do item, do BI do TAI da PB – Leitura, apresentados na Tabela 11, consideraram a escala transformada da PB – Leitura (BRASIL, 2011a), com média 500 e desvio padrão 100. É possível observar que os itens estão distribuídos pelos 5 níveis da escala de proficiência da PB – Leitura. O detalhamento dos níveis da escala será realizado na subseção 4.4.2 e a descrição dos eixos de habilidade e descritores da Matriz da PB – Leitura consta do Anexo A.

Além do ML1P, a TRI oferece outros modelos que não serão abordados neste trabalho, visto que o estudo focou a PB. Mais detalhes sobre outros modelos serão encontrados em Andrade, Tavares, Valle (2000), Muñiz (1997), Olea, Ponsoda (2003), Pasquali (2007, 2013).

O M1P pode ser representado graficamente pela curva característica do item (CCI), na qual a probabilidade de acerto ao item é função da habilidade do(a) respondente e do parâmetro de dificuldade. O gráfico plotado na Figura 18 denota que essa relação não é linear. É possível notar também que os(as) respondentes com maior habilidade possuem maior probabilidade de acertar o item. Verifica-se que a inclinação da curva em forma de “S” é dada pelo parâmetro “a” e o deslocamento na escala de habilidade é definido pelo parâmetro “b”, ambos referentes ao item. Vale citar que essa escala de habilidade é arbitrária na qual o fundamental são as relações entre os parâmetros estipulados e não necessariamente as magnitudes.

Figura 18 – Curva característica de dois itens (ML1P)



Fonte: Elaboração da autora.

Analisando em detalhe as CCI desses dois itens, segundo o modelo de M1P, verifica-se que o item 1 é considerado mais fácil que o item 2, porque o parâmetro de dificuldade ou proficiência do item “b” dele é menor. No item 1, observa-se que b = 0,4 e no item 2, b = 1,5 (valores dados na escala padronizada com média 0 e desvio 1). A dificuldade de um item indica, portanto, o grau em que a CCI está deslocada para a esquerda ou direita, sendo que um deslocamento à esquerda indica a tendência de o item ser mais fácil e o deslocamento à direita, mais difícil. Em outras palavras, quanto maior o valor do parâmetro “b”, mais difícil é o item.

Nessa CCI, outro detalhe é revelado: se a proficiência do respondente (θ) coincide com a dificuldade do item, a probabilidade de acertar $P(X = 1|\,\theta)$ = 0,5. Em outras palavras, a dificuldade de um item é o valor de θ cuja $P(X = 1|\,\theta)$ = 0,5. A função também indica que, para os itens ilustrados, se o θ é inferior ao valor – 2, a probabilidade de acerto é quase zero, significando que o modelo não considera a produção de acertos por escolha aleatória.

No M1P, a inclinação da CCI para todos os itens é a mesma, ou seja, todos os itens têm a mesma discriminação, de forma geral esse valor é 1, mas para os itens da PB o valor estimado foi de aproximadamente 1,47 na escala padronizada, como se pode

observar no fragmento da planilha fornecida à SME/SP, mostrada na Tabela 12. Nela o parâmetro “a” é identificado por A.

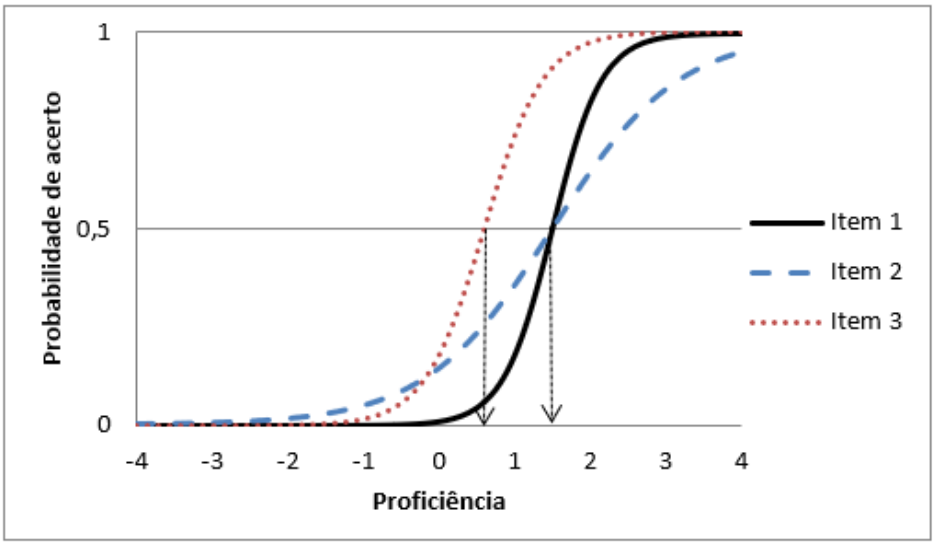
Tabela 12 – Parâmetros da TRI de 5 itens da PB – Leitura

36157	1,470	0,002	-1,729	0,099	0,000	0,000
36461	1,470	0,002	-1,834	0,105	0,000	0,000
14682	1,470	0,000	0,026	0,000	0,000	0,000
36263	1,470	0,002	-1,220	0,075	0,000	0,000
14834	1,470	0,000	-1,385	0,000	0,000	0,000

Fonte: Excerto da planilha fornecida pelo Inep à SME/SP.

O parâmetro de discriminação “a” corresponde à inclinação da curva ou ao poder de discriminação do item. O valor de “a” indica a quantidade de informação sobre a estimativa da proficiência que o item é capaz de fornecer. Quando o valor desse parâmetro é variável, essa informação interfere na determinação do erro de medida da estimativa da proficiência. A título de ilustração, na Figura 19, encontram-se as CCI de três itens, em que um deles, o item 2, apresenta parâmetro “a” diferente.

Figura 19 – Curva característica de itens com parâmetro de discriminação diferentes



Fonte: Elaboração da autora.

Para adquirir relevância e auxiliar a prática avaliativa, a escala que foi arbitrada em média zero e desvio padrão unitário, requer um processo posterior que permite ancorar itens que possam servir de referência para a interpretação pedagógica. A

escala da PB – Leitura passou por esse processo e foram definidos cinco níveis de proficiência, conforme detalhado na seção a seguir.

4.4.2 A definição dos níveis da escala de proficiência da PB – Leitura

A PB – Leitura é um teste construído segundo pressupostos da medida educacional e a estimação da proficiência é essa medida. O desenvolvimento cognitivo do(a) aluno(a), assim como outros atributos psicológicos, não é diretamente observável, diferentemente de atributos como o comprimento e a superfície. Consequentemente, o processo de medida na educação, tal como em outras ciências, passou por desafios na constituição de uma medida verdadeira que fundamentasse o estabelecimento de uma relação entre o empírico (observável) e o fenômeno que se quer aferir.

De forma sucinta, a primeira resposta ao desafio foi a Teoria Representacional da Medida (TRM), também chamada de representacionismo clássico (BORSBOOM, 2003, 2005), que embora tenha permitido a mensuração de atributos psicológicos, levou à crença de que medir é inserir números para representar uma variação a partir de alguma regra, ainda que ela não seja válida para representar as magnitudes do atributo medido, aspecto mais detalhado em Golino e Gomes (2015, p. 21-24).

Posteriormente, Krantz et al. (1971) propõem a Teoria da Medida Aditiva Conjunta (TMAC), um sistema formal de relações matemáticas que levam à medida verdadeira sem a necessidade da concatenação empírica, argumentando que isso é amplamente trabalhado também no estudo de fenômenos e objetos da Física. A ideia principal dessa teoria está na relação entre três variáveis, sendo duas independentes e uma dependente. A medida é obrigatoriamente definida pela relação simultânea entre as três variáveis, o que serviu para que fosse cunhada a expressão *medida conjunta* (BORSBOOM, 2005; KRANTZ et al., 1971). Nela, o ordenamento ocorre pela relação conjunta de duas variáveis, por exemplo, variável habilidade de pessoas (θ) e variável dificuldade dos itens (b). O ordenamento das variáveis só será possível com a fixação de uma delas, ou seja, fixada a dificuldade do item, pode-se ordenar as habilidades e vice-versa. Dessa forma, a medida passa a depender das duas variáveis simultaneamente e deixa de depender da observação direta, realizada pelos sentidos humanos, para que a concatenação seja possível.

O M1P da TRI oferece a função com propriedades que permitem mapear as relações entre os parâmetros de proficiência de pessoas e dificuldades dos itens em um sistema representacional numérico. Nesse sistema, atribuímos números a cada duas combinações de pessoas e itens e a relação de ordem está mantida no resultado dessa função que é dada pela probabilidade de acerto ou erro a um determinado item condicionada à habilidade ou proficiência.

Nessa acepção, a relação simultânea entre habilidade (proficiência) e dificuldade do item resolve o problema da impossibilidade de processos de medição considerados verdadeiros não somente nas ciências humanas e sociais. Contudo, a concatenação dos itens e das proficiências, estabelecida pela função, precisa ser interpretada para além dos aspectos numéricos. Nesse sentido, a necessidade de compreensão dos aspectos numéricos das variáveis (DE AYALA, 2009) se traduz na interpretação pedagógica da escala de medida, pois significa o quanto a representação numérica está relacionada às diferentes manifestações do atributo medido.

De modo geral, a escala de proficiência nos modelos da TRI estabelece um *continuum* de valores numéricos em ordem crescente e cumulativa, na qual são posicionados tanto os itens do teste como os(as) respondentes, tendo em vista o domínio da área de conhecimento aferida (UBRIACO, 2012, p. 88).

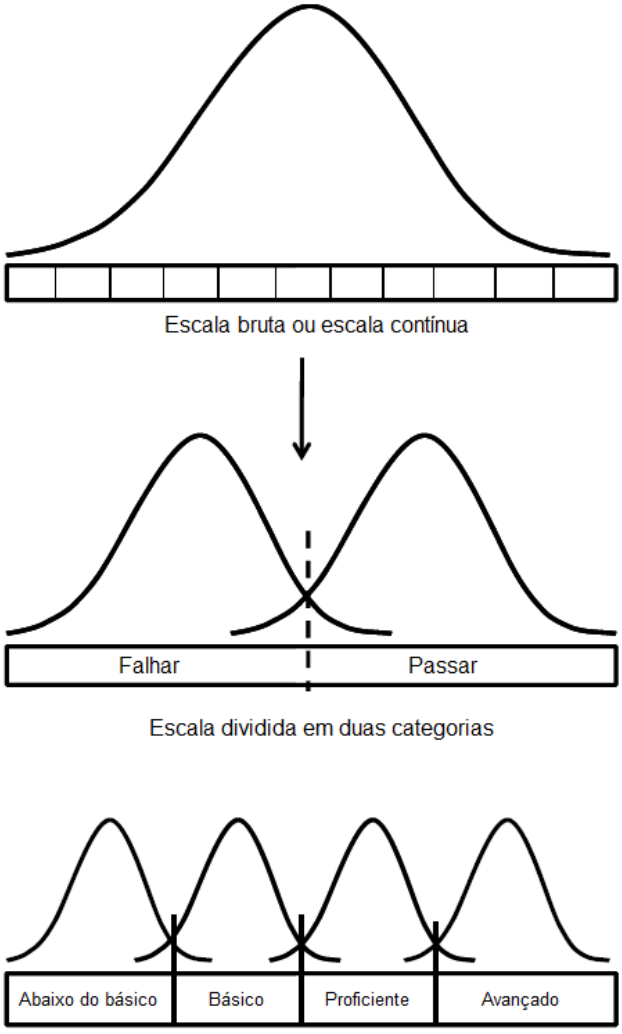
O documento Brasil (2016a) forneceu a interpretação da escala com o objetivo de proporcionar dados para que as equipes gestoras das escolas e professores(as) possam analisar e refletir sobre os resultados dos(as) alunos(as), efetivando o processo de avaliação com base na medida, conforme indicam Mujika e Etxebarría (2009).

A interpretação pedagógica, diferentemente dos resultados numéricos, proporciona critério para apoiar o julgamento e consiste em dividir o *continuum* da escala em partes, ou os chamados pontos de corte nas escalas (Cf. CISEK; BUNCH, 2007). Essa divisão e a localização de itens característicos permitem a interpretação pedagógica ou a definição dos domínios demonstrados pelos(as) respondentes em termos do traço latente aferido.

A definição do(s) ponto(s) de corte para a interpretação pedagógica dos itens recupera a relação entre a medida (número) e o atributo (proficiência em Leitura). Embora esse procedimento ainda não consubstancie a avaliação, será fundamental para que o

julgamento, essencial ao processo avaliativo, possa ser realizado. A Figura 20 ilustra diferentes formas de definir os pontos de corte em escalas.

Figura 20 – Configuração dos pontos de corte em escalas



Fonte: Zhu (2013).

A construção da escala começa pela atribuição dos parâmetros dos itens e dos respondentes por um modelo da TRI. Vale esclarecer que os valores estimados para a proficiência dos(as) respondentes e dos parâmetros dos itens, obtidos pelos *softwares* de análise pela TRI, utilizam uma escala em que são arbitradas a origem (em zero) e as unidades para θ. No caso, são arbitrados a origem com média em zero e o desvio padrão unitário (BRASIL, 2011a). Nessa representação, os valores de proficiência assumem valores negativos e nulos, os quais podem dificultar a compreensão dos resultados por parte de gestores(as), professores(as), responsáveis e alunos(as). Com efeito, pode-se escolher outra origem e outras unidades para θ.

Conforme aponta Brasil [2011b, p. 6], utilizou-se a escala com média igual a 500 e o desvio-padrão igual a 100. A decisão pela escala de desvio padrão igual a 100 se justifica pela existência de maior número de itens âncora em níveis não muito próximos entre si.

Sendo assim, os psicometristas fizeram uma transformação linear dos parâmetros dos itens e das proficiências. Essa transformação é admitida porque mantém o mesmo valor de P(θ), ou seja, P(θ') = P(θ) e oferece aos usuários resultados com valores mais compreensíveis.

Na PB – Leitura, a transformação realizada admitiu a média 500 e o desvio padrão 100, constantes de transformação fornecidas pelo Inep. Exemplificando para um item da PB – Leitura cujos parâmetros são a = 1,47 e de b = – 1,432, os valores transformados ficariam:

a' = (1,47) /100 = 0,00147

b' = 100 (– 1,432) + 500 = 356,8

Embora os resultados da PB – Leitura impressa sejam apresentados por números de acertos, para o TAI da PB – Leitura os resultados são os θ estimados e calculados na escala transformada. No teste impresso, os(as) professores(as) precisavam verificar as respostas que estavam corretas, somar os acertos e compará-los com os valores de uma tabela, fornecida nos documentos que acompanhavam o teste. Com a tabela era possível alocar os(as) alunos(as) nos cinco níveis de desempenho.

Para cada edição do teste, os psicometristas do Inep construíam uma tabela. A Tabela 13, por exemplo, mostra como era essa relação no teste 2 da edição de 2015 da PB – Leitura. A construção do TAI da PB – Leitura elimina a necessidade de a escola usar a tabela para realizar a transformação dos acertos em níveis de desempenho.

Tabela 13 – Níveis de desempenho na PB – Leitura, teste 2, edição 2015, por número de acertos no teste

Nível de desempenho	Total de acertos
Nível 1	até 3
Nível 2	de 4 a 7
Nível 3	de 8 a 12
Nível 4	de 13 a 15
Nível 5	de 16 a 20

Fonte: Brasil (2015b, p. 24).

Segundo o relatório técnico fornecido pelo Inep (BRASIL, [2011b]), a construção da escala se baseou na construção da escala do Exame Nacional para Certificação de Competências de Jovens e Adultos (Encceja), edição 2006. Como os resultados das proficiências dos(as) respondentes dos municípios não retornam para o Inep[50], os níveis de desempenho da PB – Leitura decorrem da escala de proficiência única resultante dos pré-testes realizados a cada ano.

No relatório, foi explicitado que, embora os itens pré-testados e os grupos de respondentes sejam diferentes a cada ano, a TRI permite alocação dos parâmetros dos itens e dos sujeitos em uma única escala. Isso se torna possível desde que itens calibrados, cujos parâmetros tenham sido anteriormente definidos, sejam inseridos nos pré-testes com novos itens, cujos parâmetros serão definidos (Cf. ANDRADE; TAVARES; VALLE, 2000; COUTO; PRIMI, 2011; MUÑIZ, 1997). Constrói-se assim uma escala em que respondentes e itens podem ser localizados. Com a transformação linear arbitrada pelos psicometristas responsáveis, a escala de proficiência da PB – Leitura ficou centrada no valor 500 com intervalos de 100 em 100. Considerando θ_i a representação das proficiências dos(as) i respondentes de um teste e b_j para a representação da dificuldade dos j itens alocados na escala mediante o pré-teste, pode-se ilustrar a distribuição de respondentes e itens na escala da PB – Leitura, como mostrado na Figura 21.

Figura 21 – Representação de respondentes e itens na escala da PB – Leitura

Fonte: Brasil, [2011b].

[50] A partir de 2012, foi implementada uma plataforma para inserção dos dados da Provinha Brasil.

Essa característica permitiu informar a localização de um(a) respondente ou a distribuição de respondentes de uma turma ou unidade educacional com base no resultado da aferição da proficiência pela TRI, procedimento utilizado no TAI da PB – Leitura.

A escala única, conceito essencial dos modelos da TRI, recupera para a representação matemática as relações empiricamente observáveis do desempenho das pessoas. Borsboom (2005, p. 4, tradução nossa) apresenta o seguinte exemplo para ilustrar a escala: de forma empírica João resolveu com sucesso os itens 1, 2 e 3 em um teste, enquanto Ana resolveu os itens 1 e 2, mas falhou no item 3. A representação matemática construída para essa relação pode atribuir a João um maior número do que a Ana, o que indica que ele resolveu mais itens, e atribuir ao item 3 um número maior do que os atribuídos aos itens 1 e 2, o que indica que ele foi menos frequentemente resolvido.

Essa ilustração serve para realçar a necessidade de recuperar a relação empírica que permite atribuir os itens que mais (ou menos) respondentes com determinados θ conseguem responder, e o procedimento que garantirá essa relação é dado pelos procedimentos de interpretação da escala.

Existem vários tipos de procedimentos de interpretação da escala e definição de níveis. Os cinco níveis da PB – Leitura foram estabelecidos a partir da definição dos níveis e itens âncora com a interpretação pedagógica feita por um painel de especialistas da área. As interpretações pedagógicas dos níveis de proficiência estão baseadas na definição de níveis e itens âncora.

A definição clássica de níveis e itens âncora é apresentada por Andrade, Tavares e Valle (2000, p. 110):

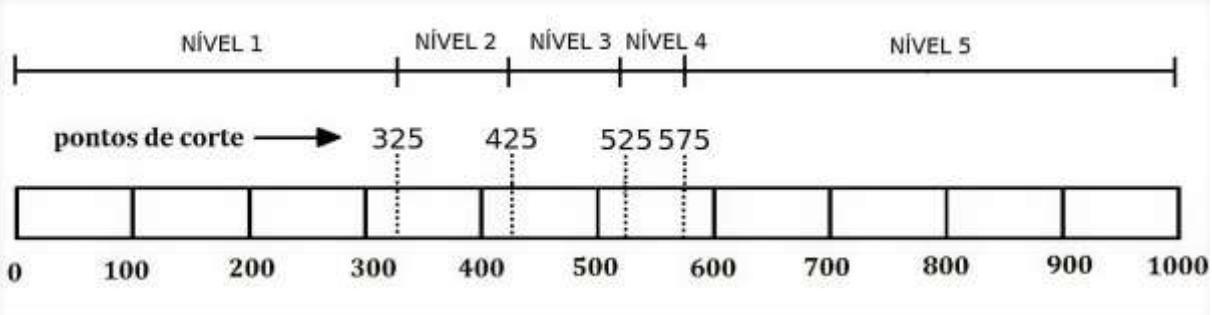
> Considere dois níveis âncora consecutivos Y e Z com Y < Z. Dizemos que um determinado item é âncora para o nível Z se e somente se as 3 condições abaixo forem satisfeitas simultaneamente:
> 1. $P(U = 1|\theta = Z) \geq 0,65$ e
> 2. $P(U = 1|\theta = Y) < 0,50$ e
> 3. $P(U = 1|\theta = Z) - P(U = 1|\theta = Y) \geq 0,30$

Na explicação dos autores, para um item ser considerado âncora em um determinado nível âncora Z da escala, ele precisa ser respondido corretamente por uma grande proporção de indivíduos deste nível âncora (pelo menos 0,65) e ser respondido por

uma proporção menor de indivíduos no nível de habilidade imediatamente anterior Y (no máximo 0,50). Adicionalmente, a diferença entre a proporção de indivíduos que responde corretamente ao item nesses dois níveis de habilidade deve ser de pelo menos 0,30.

Klein (2003, p. 138) explica que esse procedimento apresentou alguns inconvenientes, entre eles, a necessidade de maior espaçamento entre os níveis para permitir ancoragem de um número maior de itens e tornar a interpretação mais rica, além do fato de alguns descritores da matriz de referência não aparecerem na descrição do nível, mesmo quando contemplados por alguns itens, devido a não se caracterizarem como âncora. Em vista disso, o autor propôs ampliar o conceito de nível âncora para quase âncora, aumentando o conjunto de itens considerados âncora e diminuindo as distâncias entre os níveis. Os cortes da escada de proficiência em leitura da PB são ilustrados a seguir.

Figura 22 – Cortes e níveis da escala da PB – Leitura



Fonte: Elaboração da autora, com base em São Paulo (2016c).

O relatório Brasil [2011b] esclarece que a escala da PB – Leitura caracteriza os itens e níveis âncora com base na proposta de Klein, ou seja, o mesmo procedimento que fundamentou a metodologia usada no Saeb de 1999 (Cf. KLEIN, 2003). Esclareceu ainda que, após a definição de itens e níveis quase âncora, os especialistas da área de conhecimento aferida confirmaram se o corte estabelecido realmente determinava um domínio cognitivo diferente bem como interpretaram cada nível, em termos das características cognitivas exigidas nas tarefas propostas pelos itens que o representava – interpretação pedagógica do nível. Com a evidência da equivalência dos parâmetros dos itens na prova impressa e informatizada, foi possível utilizar a escala e a interpretação pedagógica do teste impresso da PB – Leitura no TAI da PB – Leitura.

4.4.3 Método de estimação da proficiência e critério de seleção de itens no TAI da PB – Leitura

O foco da primeira etapa do algoritmo TAI da PB – Leitura, núcleo central da adaptação do teste, colocou-se sobre o procedimento de estimação da proficiência.

Com um banco de itens completamente parametrizado por um modelo da TRI, é possível estimar o traço latente (θ), utilizando diferentes procedimentos. Em São Paulo (2016c) consta que foram comparados quatro métodos de estimação, os quais se desdobraram em sete, devido às variações de operacionalização e do *software* utilizado. Os métodos foram: máxima verossimilhança (ML), verossimilhança ponderada, distribuição esperada *a posteriori* (EAP)[51] e o estimador modal (BIRNBAUM, 1969; BOCK; MISLEVY, 1982; EMBRETSON; REISE, 2000; LORD, 1980; MOREIRA JUNIOR, 2011; WARM, 1989). Os dois últimos métodos se baseiam na estatística bayesiana.

A estimação da proficiência no TAI da PB – Leitura foi realizada por pacotes do repositório do R capazes de estimar proficiência a partir de parâmetros fixos de modelos da TRI. Isto foi necessário porque o pacote mais utilizado para estimativas da TRI é o ltm (RIZOPOULOS, 2006), porém ele não aceita que os parâmetros dos itens sejam fixados. Os pacotes R que satisfazem esta condição são: catR (MAGIS; RAÎCHE, 2012), PP (REIF, 2014) e irtoys (PARTCHEV, 2016). A lista dos quatro métodos com as variações de operacionalização e s*oftware* testados foram: máxima verossimilhança (ML), do pacote catR; verossimilhança ponderada (WL), do pacote catR; estimador bayesiano modal (BM), do pacote catR; método EAP (EAP), função thetaEst, do pacote catR; método EAP (eapC), função eapEst, do pacote catR; método EAP (eapI), do pacote irtoys; e método EAP (eapP), do pacote PP.

As simulações possibilitaram a comparação dos métodos quanto à menor variação do erro padrão de medida, a melhor taxa de acerto, a maior velocidade de processamento e a estimação da proficiência no intervalo da escala. O método que apresentou o conjunto de indicadores mais favorável foi o eapI, que se trata do método EAP do pacote irtoys. Foram definidos, após alguns testes, 21 pontos de quadratura para uma estimação suficientemente rápida e precisa da proficiência.

---

[51] Do termo em inglês *Expected a Posteriori*.

Bock e Mislevy (1982) apontaram as vantagens dos estimadores EAP em testes adaptativos computadorizados, nos quais se busca a eficiência da computação. As estimativas de EAP calculadas por quadratura requerem significativamente menos operações que as estimações por MAP ou por ML. As probabilidades do log empregadas nesses cálculos acumulam-se como somas simples, à medida que itens são apresentados sucessivamente. Além disso, as probabilidades de resposta nos pontos de quadratura atribuídos podem ser avaliadas antecipadamente e podem ser armazenadas com o respectivo item no BI.

A estimação da proficiência por EAP, consiste em assumir a distribuição *a priori* para os parâmetros θ, construir uma nova função denominada distribuição *a posteriori* e estimar os θ com base na média da distribuição *a posteriori*. Na estimação EAP, é fornecida uma estimativa da distribuição das proficiências dos(as) respondentes na forma de uma distribuição discreta, possível mediante acumulação das densidades de todos os sujeitos nos pontos de quadratura.

A estimação do θ, desde que sejam conhecidos os valores dos parâmetros dos itens, pode ser obtida por máxima verossimilhança (ML), que consiste em buscar o valor de θ associado ao máximo valor da função de verossimilhança. A função de verossimilhança assume que, para uma certa proficiência, a probabilidade de emitir um certo padrão de respostas é igual ao produto das probabilidades de emissão de respostas a cada item, desde que esteja satisfeito o pressuposto de independência local (Cf. ANDRADE; TAVARES; VALLE, 2000; KLEIN, 2003; MUÑIZ, 1997; OLEA; PONSODA, 2003; PASQUALI, 2013).

A função de verossimilhança é dada pela expressão:

Equação 2 – Função de verossimilhança

$$L(u|\theta) = \prod_{j=1}^{n} P_j^{u_j} Q_j^{1-u_j}$$

Em que:

u é o vetor de respostas (1 para acertos e 0 para erros)

P_j é a probabilidade de acertar o item dado um nível de proficiência θ;

Q_j é a probabilidade de errar o item dado um nível de proficiência θ.

O valor de θj que maximiza a função verossimilhança configura a proficiência estimada.

Embora a estimação da proficiência por ML seja muito mais complexa, visto que é preciso considerar o alto número de respondentes, itens, muito mais do que apenas alguns pontos para os níveis de proficiência e cálculos mais complexos, pretende-se exemplificar a estimação do θ, utilizando um teste impresso da PB – Leitura, que apresenta o mesmo conjunto de 20 itens para todos(as) os(as) respondentes, sendo que todos possuem os parâmetros.

Essa exemplificação baseou-se em Olea e Ponsoda (2003, p. 23). Será considerado um(a) aluno(a) X submetido(a) ao teste com 20 itens, cujo vetor de respostas é dado na Figura 23.

Figura 23 – Vetor X de respostas de um(a) respondente

	Respostas dadas nos Itens																				
Respondente	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
X	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	05

Fonte: A autora, com base em dados fornecidos pelo Núcleo Técnico de Avaliação da SME/SP.

Como todos os itens desse teste receberam parâmetros da TRI, pode-se calcular P(θ), usando a Equação 2 para o item acertado ou o $Q(\theta)$ para o item errado. $Q(\theta)$ é obtido por 1 – P(θ). O cálculo da probabilidade, ou seu complemento, foi realizado para valores de θ pertencentes ao intervalo [– 4 a + 4] e para todos os itens do teste. Esse cálculo, realizado por programas computacionais, foi sinteticamente ilustrado na Tabela 14 somente com alguns pontos do intervalo para θ.

O cálculo para o(a) respondente que acerta o primeiro item do teste, cujo parâmetro a = 1,47, o parâmetro b = – 1,581 e o parâmetro θ = – 3. A probabilidade de acerto ao item é condicionada à proficiência. Usa-se a função apresentada na Equação 1 para calcular a probabilidade de acerto a esse item:

$$P(-3) = \frac{1}{1 + 2{,}718^{-1{,}47.(-3-(-1{,}581))}} =$$

$$= P(-3) = \frac{1}{1 + 2{,}718^{-1{,}47.(-1{,}419)}} =$$

$$= P(-3) = \frac{1}{1 + 2{,}718^{2{,}08593}} =$$

$= P(-3) = \frac{1}{1 + 8{,}050335} =$

$= P(-3) = \frac{1}{9{,}050335} = 0{,}110493$

De modo geral, o modelo da TRI utilizado permitiu interpretar que a probabilidade P(θ) de respondentes com $\theta = -3$ acertarem o item com parâmetro b = – 1,581 é de 0,110493 e a probabilidade de respondentes com mesmo valor de θ errarem o item seria o valor complementar $Q(\theta)$, que em termos de valores será 1 – 0,110493 = 0,889506.

Conhecido o vetor de respostas do(a) respondente e os parâmetros de todos os itens do teste, podem ser calculadas as probabilidades P(θ) para os acertos ou $Q(\theta)$ para os erros para os *thetas* do intervalo [– 4 a + 4]. As probabilidades dos 20 item foram multiplicadas $\prod$ para cada valor de *theta* e representadas na coluna final da Tabela 14. A proficiência estimada corresponde ao valor máximo encontrado para esses produtos. A proficiência estimada para o(a) respondente X é indicada por $\hat{\theta}$ e tem o valor – 1,5, na escala padronizada.

Tabela 14 – Ilustração da determinação da proficiência em um teste com 20 itens, para o(a) respondente X.

Vetor	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$\prod$
q	P(θ)	P(θ)	Q(θ)	P(θ)	P(θ)	P(θ)	Q(θ)	Q(θ)	Q(θ)	Q(θ)	Q(θ)	Q(θ)	Q(θ)	Q(θ)	Q(θ)	Q(θ)	Q(θ)	Q(θ)	Q(θ)	Q(θ)	
-4	0,028	0,033	0,988	0,983	0,020	0,011	0,994	0,992	0,997	0,998	0,997	0,998	0,996	0,998	0,995	0,999	0,999	1,000	1,000	0,999	0,000
-3,5	0,056	0,066	0,975	0,965	0,042	0,022	0,988	0,984	0,993	0,995	0,994	0,995	0,991	0,996	0,989	0,997	0,999	0,999	0,999	0,998	0,000
-3	0,110	0,128	0,949	0,929	0,083	0,045	0,975	0,967	0,986	0,991	0,988	0,990	0,982	0,991	0,977	0,994	0,997	0,998	0,998	0,996	0,000
-2,5	0,206	0,234	0,900	0,863	0,159	0,090	0,949	0,934	0,972	0,981	0,975	0,979	0,962	0,981	0,953	0,987	0,995	0,996	0,997	0,991	0,000
-2	0,351	0,389	0,811	0,751	0,283	0,171	0,898	0,872	0,944	0,961	0,950	0,957	0,924	0,961	0,907	0,973	0,989	0,991	0,993	0,982	0,002
-1,5	0,530	0,570	0,674	0,591	0,451	0,300	0,809	0,765	0,889	0,921	0,901	0,915	0,854	0,923	0,825	0,944	0,977	0,981	0,985	0,963	0,004
-1	0,701	0,735	0,497	0,409	0,632	0,472	0,670	0,610	0,793	0,848	0,813	0,837	0,737	0,851	0,693	0,891	0,953	0,962	0,969	0,926	0,002
-0,5	0,830	0,852	0,322	0,249	0,782	0,651	0,494	0,429	0,648	0,729	0,676	0,711	0,574	0,733	0,519	0,796	0,907	0,924	0,938	0,857	0,000
0	0,911	0,923	0,185	0,137	0,882	0,796	0,319	0,265	0,469	0,563	0,500	0,542	0,392	0,568	0,341	0,652	0,824	0,854	0,879	0,741	0,000
0,5	0,955	0,962	0,098	0,071	0,940	0,890	0,183	0,147	0,298	0,382	0,324	0,362	0,236	0,387	0,199	0,473	0,692	0,737	0,776	0,579	0,000
1	0,978	0,981	0,050	0,035	0,970	0,944	0,097	0,076	0,169	0,228	0,187	0,214	0,129	0,232	0,106	0,301	0,519	0,573	0,625	0,397	0,000
1,5	0,989	0,991	0,024	0,017	0,985	0,972	0,049	0,038	0,089	0,124	0,099	0,115	0,066	0,127	0,054	0,171	0,341	0,392	0,444	0,240	0,000
2	0,995	0,996	0,012	0,008	0,993	0,987	0,024	0,019	0,045	0,064	0,050	0,059	0,033	0,065	0,027	0,090	0,199	0,236	0,277	0,132	0,000
2,5	0,998	0,998	0,006	0,004	0,997	0,994	0,012	0,009	0,022	0,032	0,025	0,029	0,016	0,032	0,013	0,045	0,106	0,129	0,155	0,068	0,000
3	0,999	0,999	0,003	0,002	0,998	0,997	0,006	0,004	0,011	0,015	0,012	0,014	0,008	0,016	0,006	0,022	0,054	0,066	0,081	0,034	0,000
3,5	0,999	1,000	0,001	0,001	0,999	0,999	0,003	0,002	0,005	0,007	0,006	0,007	0,004	0,008	0,003	0,011	0,027	0,033	0,041	0,016	0,000
3,5	0,999	1,000	0,001	0,001	0,999	0,999	0,003	0,002	0,005	0,007	0,006	0,007	0,004	0,008	0,003	0,011	0,027	0,033	0,041	0,016	0,000
4	1,000	1,000	0,001	0,000	1,000	0,999	0,001	0,001	0,002	0,004	0,003	0,003	0,002	0,004	0,001	0,005	0,013	0,016	0,020	0,008	0,000

Fonte: Elaboração da autora, com base em dados fornecidos pelo Núcleo Técnico de Avaliação da SME/SP.

No teste impresso da PB – Leitura, a proficiência estimada, dada por $\hat{\theta}$, é a mesma se o número de acertos é idêntico, independentemente de quais itens tenham sido acertados. Isso ocorre peculiarmente no modelo de um parâmetro da TRI, no qual o parâmetro “a” estimado é constante para os itens. Essa especificidade permitiu que os professores(as) tratassem os resultados dos(as) alunos(as) com base apenas no número de acertos.

Vale lembrar que existem as situações em que há a necessidade de estimar tanto os parâmetros como as proficiências dos(as) respondentes, processo que envolve a estimação conjunta, cuja solução consiste em resolver um complexo sistema de equações obtido em relação a cada parâmetro estimado para itens e respondentes. Hambleton e Swaminthan (1985) ainda indicam ser vantajoso o uso de ln L(u | θ), visto que o valor de θ que maximiza L(u | θ) é o mesmo para ln L(U | θ) e o produto pode ser expresso como soma dos logaritmos. Se forem considerados *N* pessoas e *n* itens, a expressão a ser maximizada ficaria:

Equação 3 – Função log-verossimilhança

$$ln\, L = \sum_{i=1}^{N} \sum_{j-1}^{n} [u_{ij}\, ln\, P_{ij} + (1 - u_{ij})\, ln\, Q_{ij}]$$

No TAI da PB – Leitura, não será realizada a estimação conjunta e mais detalhes sobre esse procedimento podem ser encontrados em Andrade, Tavares e Valle (2000), Hambleton e Swaminathan (1985), Muñiz (1997) e Olea e Ponsoda (2003).

Na estimação por ML, há um inconveniente, sobretudo nos TAI. São obtidos $\hat{\theta}$ com valores $-\infty$ ou $+\infty$ para os vetores de resposta homogêneos, em outras palavras, padrões de respostas totalmente certos ou totalmente errados. Em virtude da estimação parcial, que ocorre a cada item respondido no TAI, após responder os primeiros itens, é muito comum a obtenção de vetores homogêneos para a maioria dos(as) respondentes.

A metodologia bayesiana contorna esse inconveniente. Nos métodos bayesianos, são incorporadas à função de verossimilhança informações prévias sobre a distribuição dos parâmetros θ (*prior information*). O procedimento bayesiano requer a compreensão de dois conceitos básicos:

a) a distribuição *a priori*, uma distribuição hipotética de probabilidade para valores de *theta*, da qual se assume que os(as) respondentes são uma amostra aleatória (costuma-se usar a distribuição normal padrão);
b) a distribuição *a posteriori*, na qual a função de máxima verossimilhança (que nos dá a probabilidade de um vetor de respostas) é multiplicada pela função de distribuição *a priori*.

O procedimento EAP estabelece que o estimador de θ será a média da distribuição *a posteriori* de θ, $P(\theta \mid u)$ que é representada pela expressão:

Equação 4 – Estimador bayesiano para θ

$$P(\theta \mid u) = \frac{g(\theta)L(u \mid \theta)}{L(u)} \propto g(\theta)L(u \mid \theta)$$

Onde:

$g(\theta)$ é a função densidade (distribuição *a priori*) do θ;

$L(u \mid \theta)$ é a função de verossimilhança;

$L(u)$ é a verossimilhança do padrão de resposta u independente de θ.

Tendo em vista que o denominador é um valor concreto, a função $P(\theta \mid u)$ é proporcional ao produto da distribuição *a priori* e a função de verossimilhança.

Após a etapa da estimação da proficiência, a informação do teste passa a ter papel essencial nos TAI, pois será utilizada na seleção do próximo item. Ela será fundamental por possibilitar a diminuição do número de itens sem comprometimento da precisão da estimativa final da proficiência.

Nos TAI, as etapas de seleção de itens e finalização do teste são cruciais para maximizar a eficiência e produzir um teste curto e informativo para cada examinado(a). Os testes adaptativos atingem sua eficiência ao selecionar sucessivamente os itens que fornecem maior informação no nível provisoriamente estimado de habilidade do(a) examinando(a), mas operacionalmente são consideradas regras adicionais na seleção de itens, como a cobertura dos conteúdos e controle da exposição excessiva dos itens (PARSHALL et al., 2002, p. 127).

No TAI da PB – Leitura, a cada item respondido um θ provisório era estimado e, se as condições de encerramento não eram atingidas, um novo item era selecionado, seguindo a regra da Máxima Informação de Fisher (BARRADA, 2010).

Além dessa regra, tinha sido cogitado o acréscimo de uma regra relativa ao conteúdo, na qual os itens mais informativos seriam selecionados de modo a equilibrar a presença de itens referentes os eixos 1 e 2 da Matriz de Referência da PB – Leitura no teste. Esse acréscimo tinha o propósito de atribuir maior validade de conteúdo ao teste, garantindo melhor cobertura da matriz. Contudo, devido ao limitado número de itens do banco, não foi possível acrescentar os controles de conteúdo e de exposição de itens.

A seleção do item pelo critério da máxima informação de Fisher dependeu da aferição da informação do teste no θ provisoriamente estimado. Parshall et al. (2002) escreve que a seleção de item pela informação máxima durante o teste é muito simples computacionalmente, porque o cálculo mais pesado das funções de informação pode ser feito antes de qualquer examinando ser testado. Os resultados são apresentados em uma matriz, por valores discretos de proficiência. Hambleton, Jones e Rogers (1993, p. 3) mostram que um particular subconjunto de itens pode oferecer maior informação na estimativa da proficiência de determinado(a) respondente. Na TRI, conforme aponta Birnbaum (apud MUÑIZ, 1999, p. 127), a fórmula geral de informação do teste é dada pela soma das informações fornecidas por cada item.

Equação 5 – Fórmula geral de Informação do item

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)}$$

Equação 6 – Fórmula geral de Informação do Teste

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta)$$

Em que:

$I_i(\theta)$ é a informação do item;

$I(\theta)$ é a informação do teste e também um indicador de precisão do teste;

n é o número de itens;

$P_i'(\theta)$ é a derivada de $P_i\,(\theta)$;

$P_i\,(\theta)$ é o valor da probabilidade condicionada de acerto ao item;

$Q(\theta)$ é igual a 1 - $P_i\,(\theta)$.

A função de informação (FI) indica a precisão do teste e quanto maior a informação menor será o erro de medida

Tendo em vista que a I(θ) pressupõe a derivada da função de probabilidade [$P_i'(\theta)$], as funções de informação do teste I(θ) para os modelos logísticos de 1 e 2 parâmetros da TRI, relacionados à PB – Leitura podem ser especificadas conforme o Quadro 4.

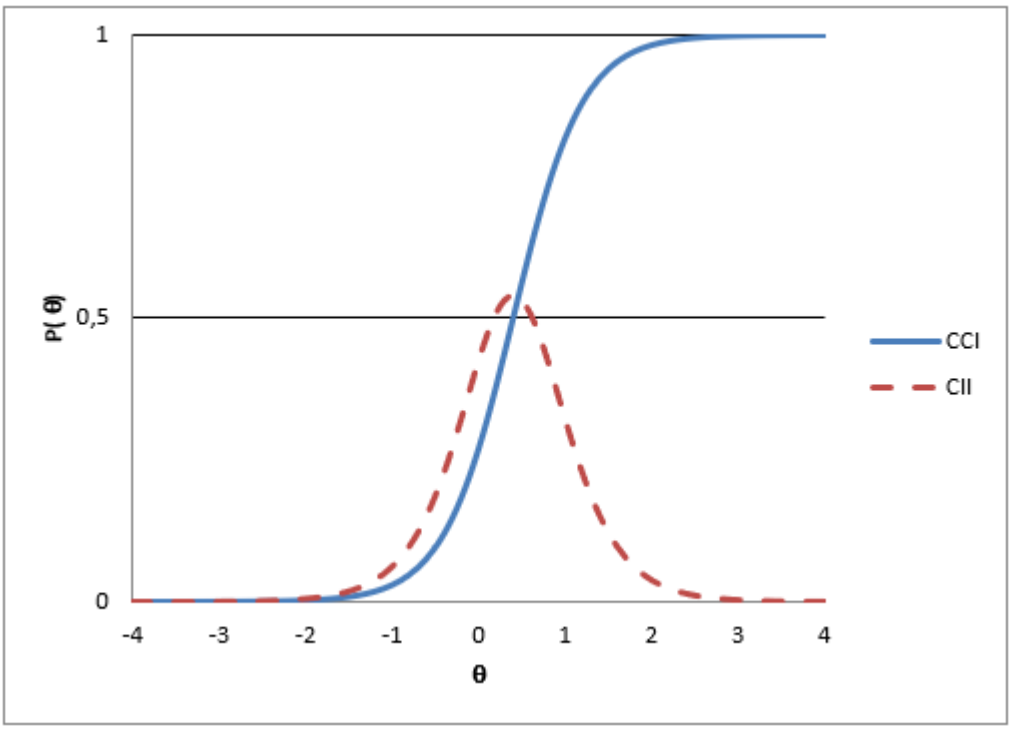
Quadro 4 – Informação do teste, por modelo da TRI

MODELO LOGÍSTICO	$\boldsymbol{P_i'(\theta)}$	**I(θ)**
1P	$DP_i(\theta)Q_i(\theta)$	$\sum_{i=1}^{n} D^2 P_i(\theta)Q_i(\theta)$
2P	$Da_iP_i(\theta)Q_i(\theta)$	$\sum_{i=1}^{n} D^2 a_i^2 P_i(\theta)Q_i(\theta)$

Fonte: Muñiz (1997, p. 127) e Hambleton e Swaminathan (1985, p. 91).

A informação do item e do teste condicionadas ao $\hat{\theta}$ podem ser descritas por curvas. A Figura 24 contém a curva de informação do item (CII) e a curva de informação de um teste (CIT) com um item, no ML1P dicotômico e unidimensional.

Figura 24 – Curva característica e de informação para um item

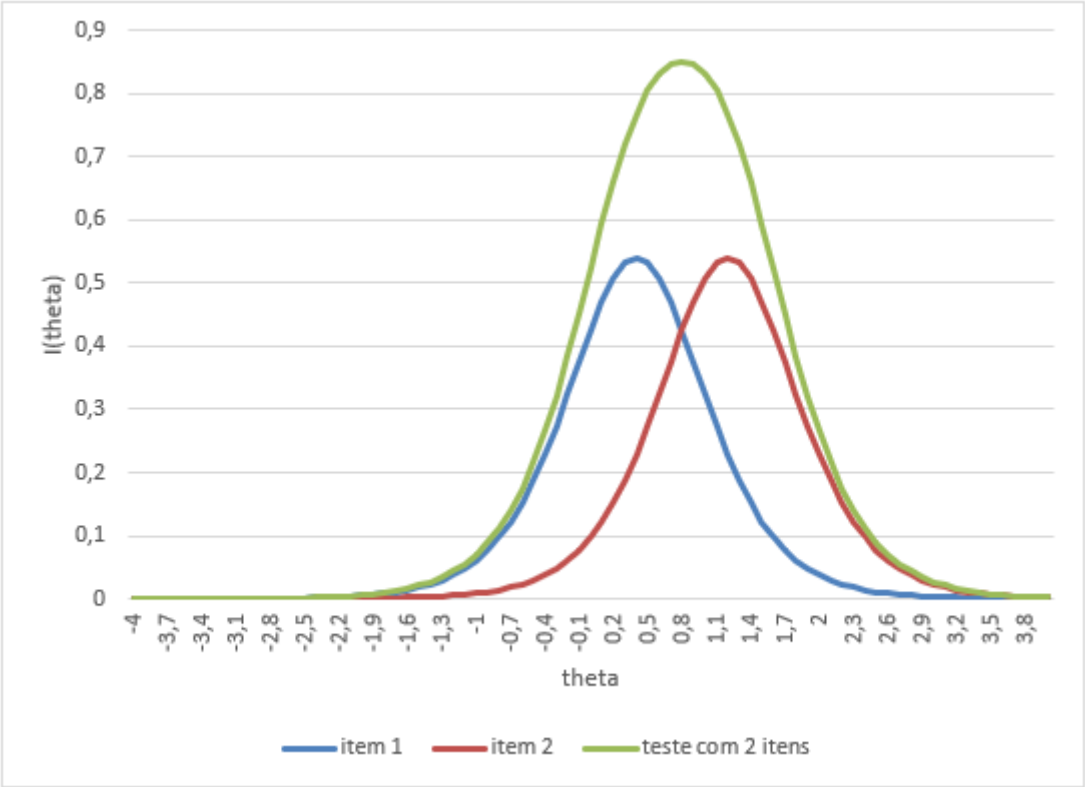


Fonte: Elaboração da autora.

As curvas de informação dos itens variam em função dos valores de θ estimados, por isso vale também chamá-las de função de informação do item (FII), sendo que a informação é máxima para determinado valor de θ. No caso do item mostrado na Figura 24, a informação máxima é obtida em θ ≅ 0,4, que é respectivamente o valor do parâmetro “b” do item. Para os ML1P, a máxima informação é obtida para $\hat{\theta}$ = b.

A Figura 25 contém a curva de informação de dois itens item (CII) e a curva de informação de um teste (CIT) com dois itens, no ML1P dicotômico e unidimensional.

Figura 25 – Curva de informação de dois itens (CII) e do teste (CIT) com dois itens



Fonte: Elaboração da autora.

Para um teste contendo esses dois itens (b= 0,4 e b = 1, 2), a maior informação é dada para os(as) respondentes com valores de θ estimados em aproximadamente 0,7.

No TAI, a estimação da proficiência a cada item respondido e a informação do teste previamente calculada para o BI possibilitam que o próximo item escolhido para compor o teste proporcione itens com alta informação para o(a) respondente e, consequentemente, um teste mais preciso.

4.4.4 O critério de encerramento do teste no TAI PB – Leitura

Na etapa de encerramento do TAI da PB – Leitura, o principal objetivo foi proporcionar um teste com menor número de itens e que agregasse a maior precisão possível. O algoritmo direcionava ao(à) respondente os itens mais informativos, ou seja, aqueles com parâmetros “b” mais semelhantes às proficiências estimadas. A cada resposta dada, o algoritmo do TAI da PB – Leitura cotejava os resultados da estimação da

proficiência e do erro de medida correspondente com as três regras do critério de finalização.

Esse critério, conforme apontado, verificava a estimação provisória da proficiência e encerraria o teste se uma das regras do critério era atingida. Considerando um TAI com objetivo de situar a proficiência do(a) respondente na escala de proficiência, as regras amplamente utilizadas na etapa de finalização do teste, conforme aponta Barrada (2012), estão pautadas em determinar um limite máximo para o erro de medida do teste ou um limite máximo e mínimo de itens do teste. Em sentido diferente, os testes com objetivo de certificação costumam se pautar na classificação da proficiência, verificando se o intervalo de confiança da proficiência estimada está em determinada subdivisão da escala, evitando que um ponto de corte possa se localizar nesse intervalo construído. No TAI da PB – Leitura, a regra dos testes com objetivo de certificação figurou como uma terceira regra. Portanto, o critério de finalização ficou constituído pelas regras:

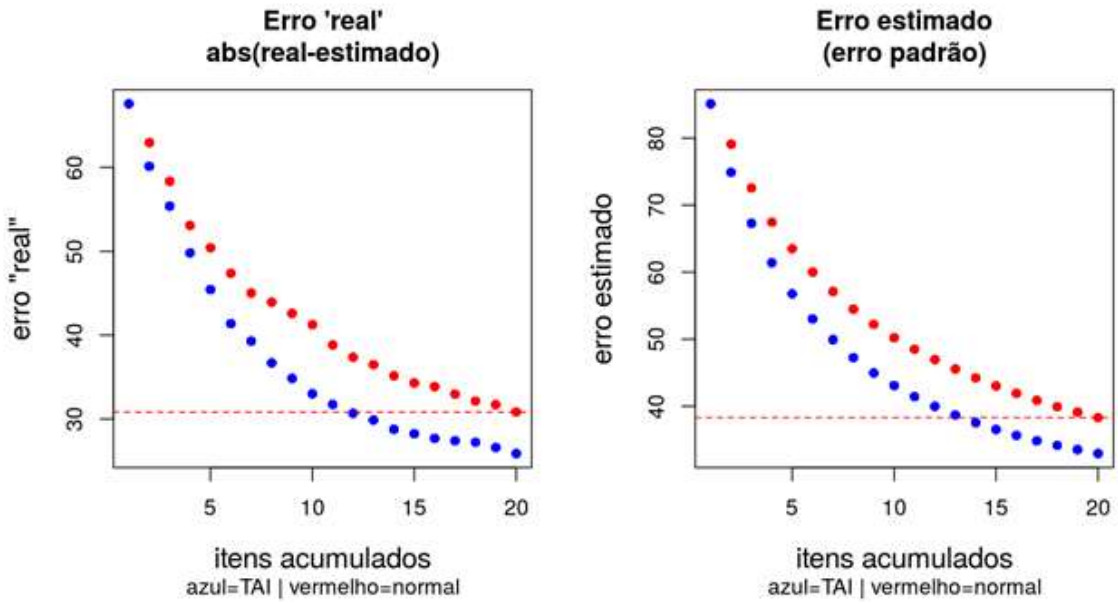
a) um número mínimo de 7 itens e máximo de 20 itens era administrado no teste;
b) o erro padrão da estimativa da proficiência, definido em 35 pontos na escala de proficiência da PB – Leitura, era atingido;
c) ter o intervalo de confiança[52] da proficiência estimada contido em um dos níveis da escala da PB – Leitura.

Para a regra (a), pautou-se na razoabilidade em produzir um teste que tivesse menor comprimento do que o teste convencional (20 itens), não adaptativo. A definição do valor de erro padrão máximo na regra (b), se baseou nas simulações (São Paulo, 2016d), mostradas na Figura 26, em que os erros de medida do teste impresso são indicados pelas linhas pontilhadas de cor vermelha e dos testes adaptativos são indicados pelas linhas pontilhadas de cor azul). O gráfico da direita ilustra o erro padrão do *theta* estimado e o gráfico da esquerda ilustra o erro padrão do *theta* verdadeiro, que só é conhecido no contexto da simulação. A linha tracejada vermelha,

[52] No desenvolvimento do algoritmo do TAI foi utilizado o intervalo de confiança, quando deveria ter sido utilizado o intervalo de confiança bayesiano (Cf. EHLERS,2007; SCACABAROZI; DINIZ, 2010).

paralela ao eixo horizontal de cada gráfico, serve de guia para o erro máximo do teste impresso com 20 itens.

Figura 26 – Erros estimados na simulação do TAI da PB – Leitura, por forma de administração (impressa e adaptativo), por tipo de estimação (theta verdadeiro e estimado) e número de itens



Fonte: São Paulo (2016c).

Com efeito, os dois gráficos revelam que um teste adaptativo, variando entre 12 e 14 itens, tende a produzir erros de medida semelhantes a um teste convencional de 20 itens. Também com fundamento nos dois gráficos, verifica-se que o erro padrão variou de 31 a 38 pontos na escala. Com base nessas observações, delimitou-se um erro padrão de 35 pontos na escala, correspondente a um teste convencional com tamanho entre 15 e 19 itens. Tal ajuste procurou o equilíbrio entre precisão e tamanho do teste em cada situação de avaliação. As semelhanças existentes nas duas simulações (*theta* estimado e verdadeiro), asseguraram que os resultados do *software* corresponderiam aos resultados da aplicação experimental. A regra de encerramento de TAI com base no limite para o erro padrão está disponível em pacotes como catR.

Cabe destacar que o propósito da PB – Leitura, preservado para o TAI da PB – Leitura, relacionou-se tanto ao objetivo de situar os(as) respondentes quanto ao de classificar com precisão se eles superam ou não um determinado nível (Cf. CIZEK; BUNCH, 2007; RENOM; DOVAL, 1999), a julgar pela existência de uma escala de proficiência

com pontos de corte e a classificação dos(as) respondentes em um dos cinco níveis dessa escala.

Nessa acepção foi incluída a regra (c) no critério de finalização. Ela foi inicialmente pensada e implantada no algoritmo sem que a ampla bibliografia sobre o assunto estivesse totalmente revisada, de modo que não tinham sido acessados os artigos que tratavam dos critérios de encerramento para testes com fins de certificação, denominados testes de classificação. Somente após a aplicação experimental e com a ampliação da revisão da bibliográfica, foram localizados artigos que corroboraram essa terceira regra (EGGEN; STRAETMANS, 2000; KINGSBURY; WEISS, 1983; MAGIS; MAHALINGAM, 2015; SPRAY; RECKASE, 1994; WEISS, 1982; WEISS; KINGSBURY, 1984). No entanto, o uso da regra de classificação, como critério de finalização, nesses artigos, vincula-se aos propósitos da certificação e substituem a regra do erro padrão máximo. De modo diferente, no TAI da PB – Leitura utilizou-se a regra de classificação adicionalmente, com o objetivo de favorecer a identificação precisa dos níveis de domínio em leitura.

O termo classificação, que denomina a regra, está associado à capacidade de colocar em classes ou categorias ou ainda identificar agrupamentos afins, definidos pelo conhecimento demonstrado no teste. Colocar em classes ou categorias por afinidade nos domínios de um dado conhecimento é extremamente relevante para a atuação pedagógica, ao contrário do uso do termo classificação na acepção de ranqueamento, que encerram em si o trabalho avaliativo. Identificar essas categorias é crucial para a definição dos domínios e consequentemente a intervenção diferenciada para cada um(a) deles(as), na perspectiva da pedagogia diferenciada (PERRENOUD, 2000).

Essa terceira regra permitiu que a administração adaptativa de itens encerrasse assim que o intervalo de confiança (de 85%) da proficiência estivesse contido em um dos níveis da escala da PB – Leitura. Desse modo a proficiência de um(a) participante do teste poderia ser classificado(a) com precisão em um nível, sem necessariamente ter um erro de medida de pequenas dimensões.

A ilustração de uma situação com essas características é mostrada na Figura 27, encontrada em São Paulo (2016d). O(a) respondente B tem seu θ estimado e respectivo intervalo de confiança, representados por 

, por sua vez, o(a) respondente A tem seu θ estimado e respectivo intervalo de confiança, representados

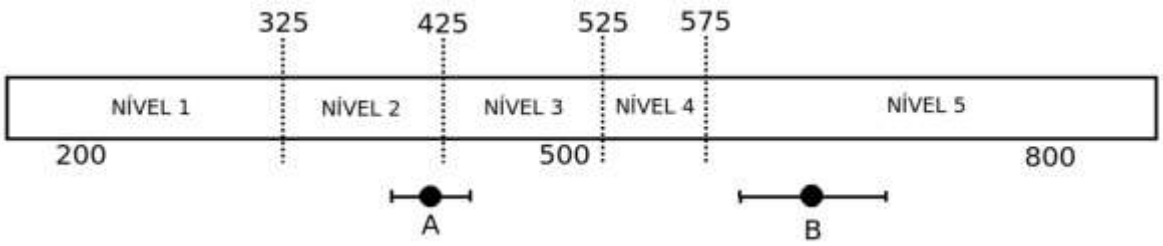
por 

. O intervalo do(a) respondente B é maior, significando erro de medida de maior dimensão do que o do(a) respondente A, mas sua proficiência já foi estimada com segurança no nível 5, depois que ele acertou 7 itens no teste adaptativo. O(a) respondente A, por sua vez, já respondeu a 11 itens no teste adaptativo, mas ainda não atingiu o critério (c), e seu intervalo encontra-se na fronteira entre os níveis 2 e 3 da escala de proficiência da PB – Leitura.

A quantidade e o posicionamento dos pontos de corte da escala da PB – Leitura interferem fortemente na interpretação pedagógica e na decisão educativa decorrente dela, e o(a) respondente B foi beneficiado pela adição dessa terceira regra ao critério de finalização de teste, terminando mais rapidamente o teste sem perder a precisão quanto ao critério de classificação em um nível da escala.

A terceira regra do critério de finalização mantém estreita a relação com a finalidade prática da PB – Leitura, proporcionando uma medição confiável da proficiência do(a) aluno(a) ao alocá-lo em um dos cinco níveis de proficiência, mesmo quando os erros de medida são maiores que os do teste impresso. Dessa forma, os(as) professores(as) podem apoiar suas decisões pedagógicas a partir de uma informação mais fidedigna. Destaca-se, sem entrar no mérito dessa decisão, que especialistas em alfabetização, consultados pelo Inep, definiram como desejável que cada aluno(a) esteja pelo menos no nível 4 ao final do 2º ano.

Figura 27 – Representação dos intervalos de confiança das proficiências verdadeiras para os(as) respondentes A e B na escala de proficiência da PB – Leitura



Fonte: São Paulo (2016c).

A definição do nível de confiança para o intervalo também se pautou nos estudos de simulação, conforme aponta relatório Brasil (2016c). Buscou-se ajustar esse nível para que pudesse abarcar uma grande quantidade de testes finalizados, sem torná-lo muito baixo. Foram testados quatro níveis de confiança.

As simulações mostraram que o aumento no nível de confiança corresponde à diminuição no número de testes finalizados.

A Tabela 15 mostra essa relação e foi decidido usar o nível de confiança de 85%, pela razoabilidade entre a taxa de acerto ao encerrar pela regra do nível de confiança e o nível de confiança.

Tabela 15 – Frequência de testes finalizados ou não pela regra de classificação na simulação do TAI da PB – Leitura, por nível de confiança

Nível de confiança	Testes finalizados com menos de 20 itens	Percentual de testes finalizados sem erro pela regra (c)
80%	354	83,9
85%	242	85,5
90%	114	95,6
95%	62	100,0

Fonte: São Paulo (2016c).

O critério de parada que, nos demais TAI, ou compõem os objetivos de classificar ou de situar respondentes, no TAI da PB – Leitura, diferentemente, priorizou-se, na estimação da proficiência, tanto na minimização do erro de medida como na abrangência do intervalo da proficiência em um único nível interpretado da escala.

A composição desse critério de parada ou finalização permitiu diferenciar uma situação que o TAI da PB – Leitura pretende resolver de forma vantajosa para o processo de avaliação de caráter formativo. Trata-se de valorizar a inserção do intervalo de confiança da proficiência verdadeira do(a) respondente inteiramente em um dos níveis da escala de proficiência. Ao corroborar para essa inserção, desconhecendo a proficiência verdadeira, estaria garantido o oferecimento de um teste para definir com maior precisão em qual nível da escala o(a) aluno(a) se encontra, auxiliando o trabalho diagnóstico dos domínios de leitura e colaborando para apoiar as ações pedagógicas decorrentes.

Para ilustrar a obtenção do intervalo de confiança usado no critério de parada, considerar-se-á um(a) aluno(a) que respondeu ao TAI da PB – Leitura e encerrou o teste após 16 itens, explicitado no Quadro 5.

Quadro 5 – Dados do respondente 46346 no TAI da PB – Leitura, submetido a um teste com 16 itens

Item	Acerto ou erro	*theta* estimado	Parâmetro b do item	Descritor	Eixo da matriz
75	0	447,5751746	500,091	D7.3	Leitura
71	0	407,1065686	448,273	D2.1	Aprop. do sist. de escrita
95	1	440,1382516	404,369	D6.1	Leitura
72	0	411,7455293	430,487	D3.4	Aprop. do sist. de escrita
67	0	386,6776161	399,288	D1.1	Aprop. do sist. de escrita
92	1	408,4612441	394,571	D5.1	Leitura
96	0	392,7024705	424,471	D2.1	Aprop. do sist. de escrita
69	0	373,4170310	363,277	D5.1	Leitura
93	0	358,2901239	372,670	D1.3	Aprop. do sist. de escrita
87	1	371,2670556	345,299	D4.1	Leitura
94	0	359,4847496	379,297	D3.5	Aprop. do sist. de escrita
65	0	346,2517011	341,897	D4.1	Leitura
90	1	358,0161261	352,965	D3.3	Aprop. do sist. de escrita
79	1	374,1954264	455,284	D10.1	Leitura
91	0	364,0363837	361,400	D3.4	Aprop. do sist. de escrita
101	1	378,1385077	457,481	D8.1	Leitura

Fonte: Elaboração da autora, com base nos microdados fornecidos pela SME/SP.

O valor de θ estimado, por exemplo, de 378,1385, considerando a origem da escala em 500 e o desvio padrão 100. Esse(a) aluno(a) foi alocado(a) no nível 2 da escala da PB – Leitura, com base na localização da proficiência estimada. Cabe observar que os itens selecionados nesse teste estão equilibrados quanto aos dois eixos da matriz, embora a programação do critério de seleção não tenha contemplado a regra de balanceamento de conteúdo.

A estimação dimensionou o erro para essa proficiência em 34,6387843. O intervalo com 85% de confiança para o θ foi estabelecido, conforme exemplificado a seguir:

NC de 85%: Z_c = 1,44

Erro =34,6387843

Erro Z_c = 1,44. 34,6387843 = 49,8798494

$\hat{\theta}$ – 49,8798494 ≤ θ ≤ $\hat{\theta}$ + 49,87984944

378,1385 – 49,87984944 ≤ θ ≤ – 378,1385 + 49,87984944

328,2586583 ≤ θ ≤ 428,0183571

Significa que foi considerado o intervalo [328,2586583; 428,0183571], com nível de confiança de 85% e, para esse respondente, não foi possível alocar totalmente o intervalo em um nível de proficiência, lembrando que os pontos de corte da escala da PB – Leitura são: 325, 425, 525 e 575. Vale retormar que o intervalo de confiança deveria considerar a distribuição *a posteriori,* para corresponder ao intervalo de confiança bayesiano.

Na situação ilustrada, o teste foi encerrado pela regra do limite para o erro padrão, estipulado em 35 pontos. Em contrapartida, se esse(a) aluno(a) fosse submetido(a) ao teste 2 da PB – Leitura, na versão impressa, embora usando outro procedimento para definição do erro de medida, cuja comparação não é apropriada, o erro de medida estipulado seria na ordem de 39,55315 e ele teria que responder a 20 itens, ao contrário dos 16 no TAI.

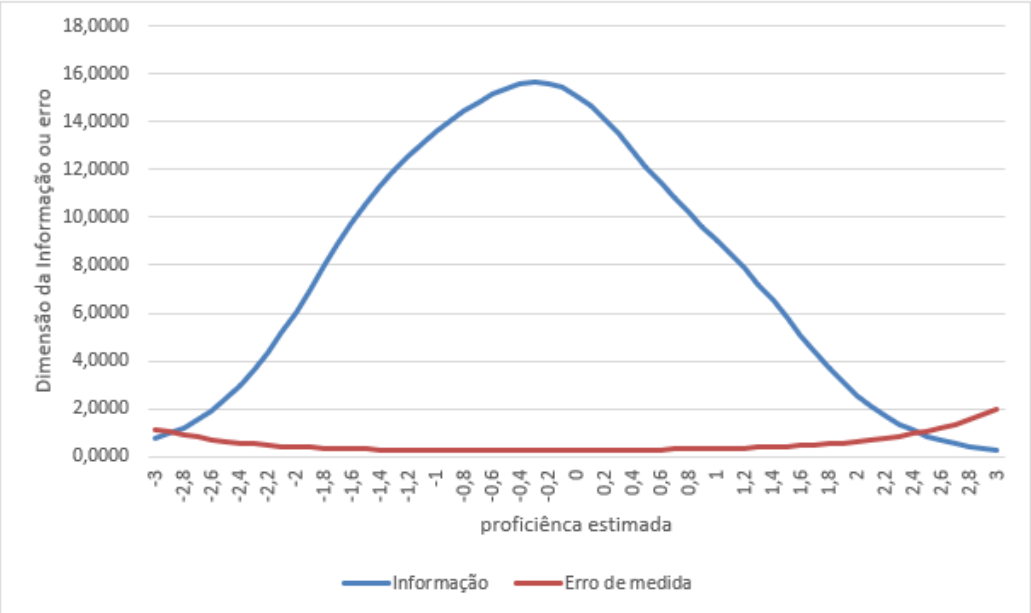
A fórmula para obtenção do erro a partir da função de informação, nesse caso, é dada por:

Equação 7 – Erro padrão do teste

$$S_e = [I(\theta)]^{-1/2}$$

Para os testes impressos, a estimação da proficiência ocorre com precisão variável, pois as estimações serão mais precisas para os $\hat{\theta}$ em que a função de informação do teste é maior ou o erro de medida é menor, de acordo com a Figura 28.

Figura 28 – Curva de informação e erro de medida na PB – Leitura, teste 2 da edição 2015

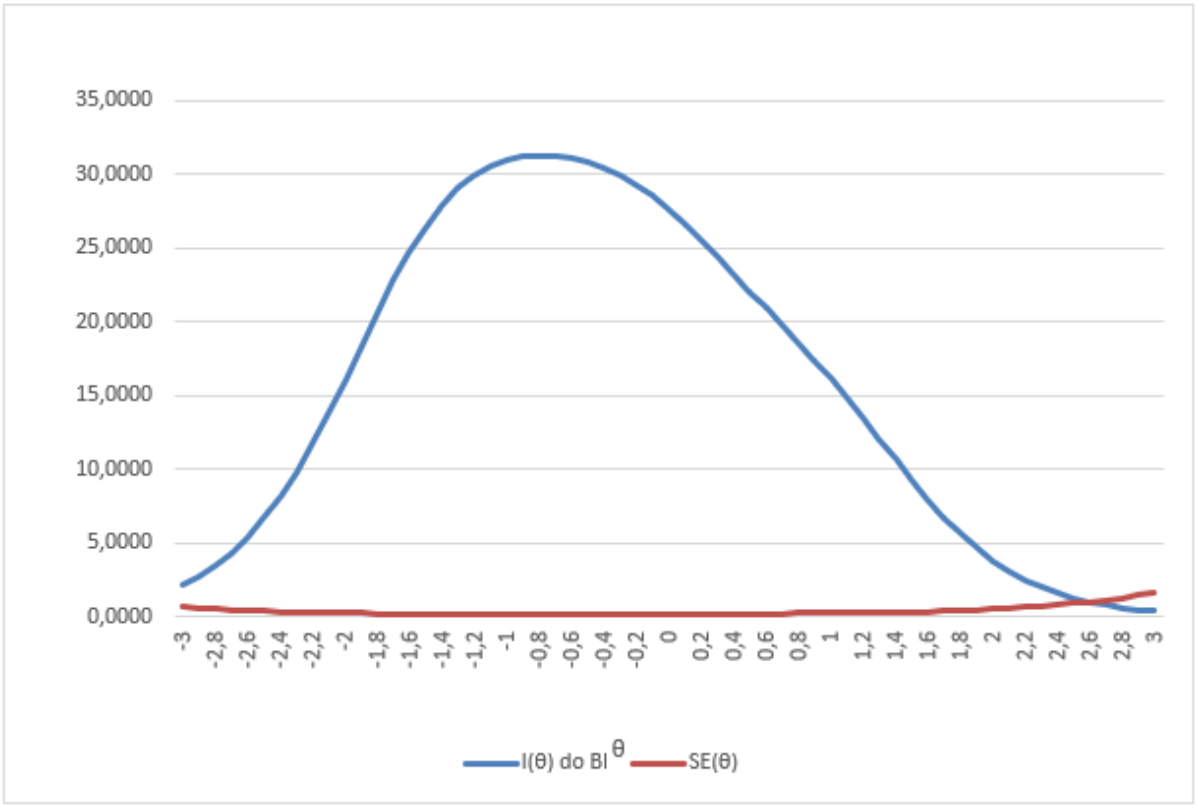


Fonte: Elaboração da autora, com base em São Paulo (2016c).

Embora nenhum respondente responda aos 39 itens do BI do TAI da PB – Leitura, a função de informação desse banco é apresentada na Figura 29.

Em um conjunto de itens com parâmetros b = { – 1,90; – 1,65; – 1,58; – 1,55; 0,00; 0,08; 0,11;0,17; 0,19; 1,05; 1,20; 1,35}, definidos na escala padronizada, observa-se que, para os(as) estudantes com baixo valor de θ, somente os quatro primeiros itens agregam informação em níveis significativos, enquanto que, para estudantes com altos valores de θ, somente os últimos itens do conjunto cumprirão a tarefa de acrescentar informação significativa à estimativa.

Figura 29 – Curva de informação e erro de medida do BI do TAI da PB – Leitura



Fonte: Elaboração da autora, com base em São Paulo (2016c).

Diante do pressuposto da TRI, para o erro de medida, pode-se concluir que: 1) um conjunto fixo de itens, situação em que o teste é denominado por linear, a informação não é maximizada para todos os níveis de proficiência dos(as) respondentes; 2) para um determinado valor de proficiência, certos itens do teste seriam dispensáveis, tendo em vista fornecerem pouca informação; e 3) o teste poderia ser reduzido aos itens que oferecem mais informação para o nível de proficiência do(a) respondente, permitindo um teste ajustado.

A conclusão de que o teste poderia ser reduzido aos itens cujas dificuldades (parâmetro b) se mantêm nas proximidades da proficiência estimada do(a) respondente apresenta um conflito nos testes impressos, pois não se sabe de antemão (antes de findar o teste) qual é a proficiência dos(as) respondentes. Um teste adaptativo informatizado responde a esse conflito, porque a cada item respondido é realizada uma estimação parcial da proficiência e, após, é selecionado o item que agrega mais informação à proficiência estimada.

Trazer o erro de medida para patamares que auxiliam a classificação dos(as) respondentes nos níveis da escala de proficiência é de fundamental importância, pois, conforme já foi apontado, a precisão tem profundas repercussões na tomada de decisão diante do processo de aprendizagem da leitura, e diminuir sua incidência será crucial para que a interpretação pedagógica tenha coerência e validade educacional (Cf. RHOADES; MADAUS, 2003; STECHER, 2002).

4.4.5 O procedimento de aplicação e resultados do TAI da PB – Leitura

Na aplicação do TAI da PB – Leitura, para evitar a possibilidade de viés em razão de problemas na locução informatizada, observada na aplicação do TBC, que só permitia uma repetição, foi aumentado o número de vezes que o(a) aluno(a) poderia repeti-la.

Inicialmente, somente os(as) alunos(as) do 2º ano participariam da aplicação experimental TAI da PB – Leitura, mas com a constatação de ausências nos dias marcados para aplicação do TBC e tendo em vista a impossibilidade de retorno para outra aplicação em cada escola, decidiu-se por envolver algumas turmas de 1º ano, considerando a suposição de que essas turmas já contemplariam conhecimentos referentes ao início do 2º ano, momento em que a PB – Leitura também é aplicada, por já estarem no final do ano letivo de 2016.

As aplicações não consideraram todas as turmas de 1º ano, devido às restrições de datas e horários para aplicação. Foi necessário realizar essa aplicação com apenas uma visita por escola e considerar que a permanência do(a) aluno(a) não poderia ultrapassar o turno de estudo, com duração de 5 horas. Com esse tempo de duração, conseguiu-se atender um limite máximo de 6 turmas por escola, com tempo médio para aplicação do TAI de 30 minutos por turma. Nesse período de tempo, privilegiou-se a aplicação para turmas de 2º ano, já que constituíam o público-alvo inicial para o

teste, avaliando as turmas de 1º ano no período restante, até o encerramento do turno escolar.

A Tabela 16 exibe o número de aplicações por escola, incluindo primeiros e segundos anos, o horário de início da primeira aplicação e o horário de fim da última aplicação. Caso uma aplicação não fosse encerrada pelo(a) aluno(a) por motivos adversos, ela era automaticamente finalizada após 2 horas. É imperativo enfatizar que, assim como em qualquer prova, aplicações incompletas do TAI podem não refletir a proficiência final real do(a) examinando(a). Para tanto, testes incompletos não foram considerados nas análises estatísticas.

Tabela 16 – Número de testes e tempo das aplicações do TAI da PB – Leitura, por escola

Escola	Número de testes aplicados	Início da aplicação	Fim da aplicação
Angola	171	09:11	17:43
África do Sul	128	14:01	16:45
Moçambique	190	10:13	11:31
Libéria	138	08:42	11:58
Argélia	93	14:29	17:31
Costa do Marfim	150	14:03	17:44
Líbia	69	14:53	17:26
Cabo Verde	154	14:13	18:03
República do Congo	148	13:54	17:46
Camarões	142	14:32	17:18
Etiópia	77	09:05	17:40
Benim	101	14:39	16:59
Ruanda	165	09:41	16:48
Egito	141	14:23	17:16
Marrocos	116	09:32	12:07
Total	1983	-	-

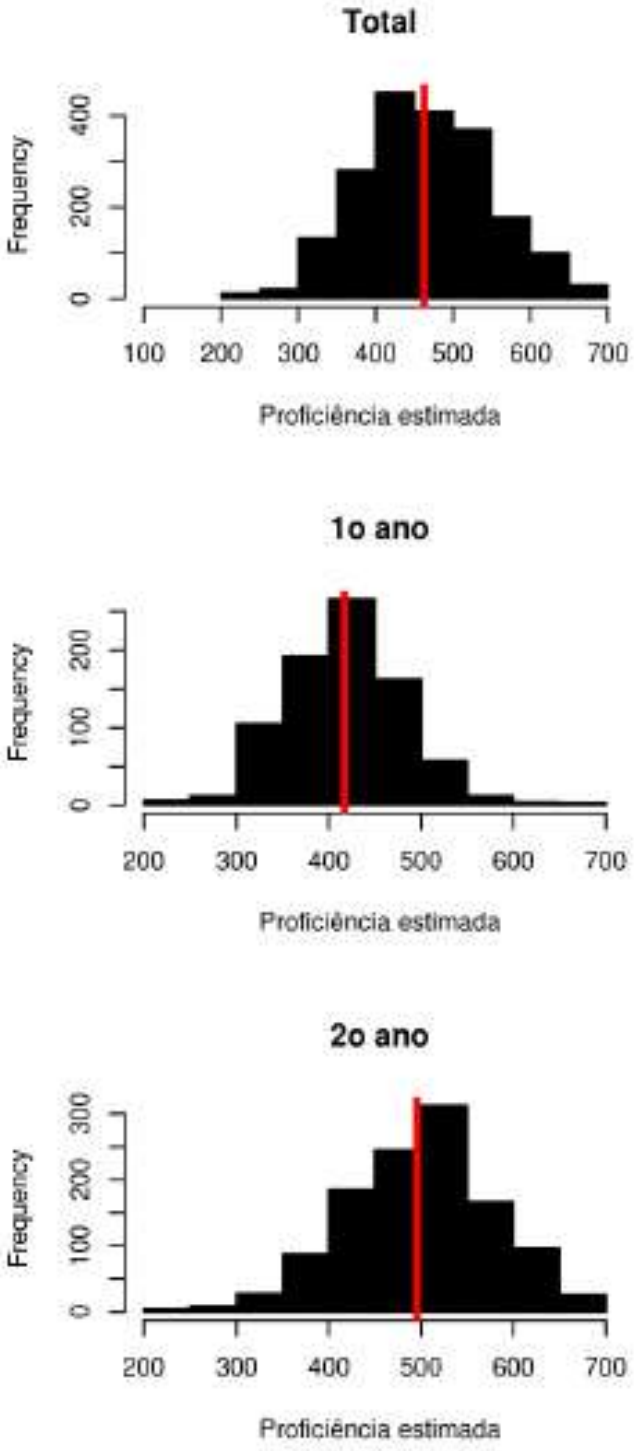
Fonte: São Paulo (2016b) com substituição do nome das escolas.

Para nortear as análises, a apresentação dos dados assentou-se nos aspectos vantajosos levantados para um TAI.

Considerando a vantagem de fornecer a pontuação imediatamente após o término da aplicação, o TAI da PB – Leitura foi aplicado para 1.983 estudantes, distribuídos em 78 turmas pertencentes às 15 escolas, sendo que os resultados eram obtidos imediatamente ao final do teste.

A partir das proficiências estimadas (Figura 30) e os pontos de corte definidos para a PB – Leitura, o algoritmo estipulou, imediatamente após o término do teste, o nível de proficiência de cada estudante.

Figura 30 – Distribuição das proficiências estimadas no TAI da PB – Leitura, com linha de média



Fonte: São Paulo (2016d).

Como esperado, os(as) alunos(as) de 1º ano se concentraram nos níveis inferiores da escala, enquanto os(as) alunos(as) de 2º ano, que são o público-alvo original tanto da PB – Leitura impressa como do TAI da PB – Leitura, tiveram seus resultados nos níveis superiores da escala (Figura 31).

A Tabela 17 e a Figura 31 também trazem o resumo dos dados. Pode-se observar que a média da proficiência para o 2º ano foi de 495,28 pontos (considerando a escala de origem em 500 e desvio padrão 100), menor que a observada para o 1º ano, validando as metodologias subjacentes à construção do algoritmo, a elaboração dos itens e a adequação da plataforma.

Essas informações mostram que o TAI da PB – Leitura também é uma plataforma flexível, que pode ser aplicada a essas diferentes populações sem perda de confiabilidade e consistência no nível das escolas. As médias das escolas estão no Anexo C.

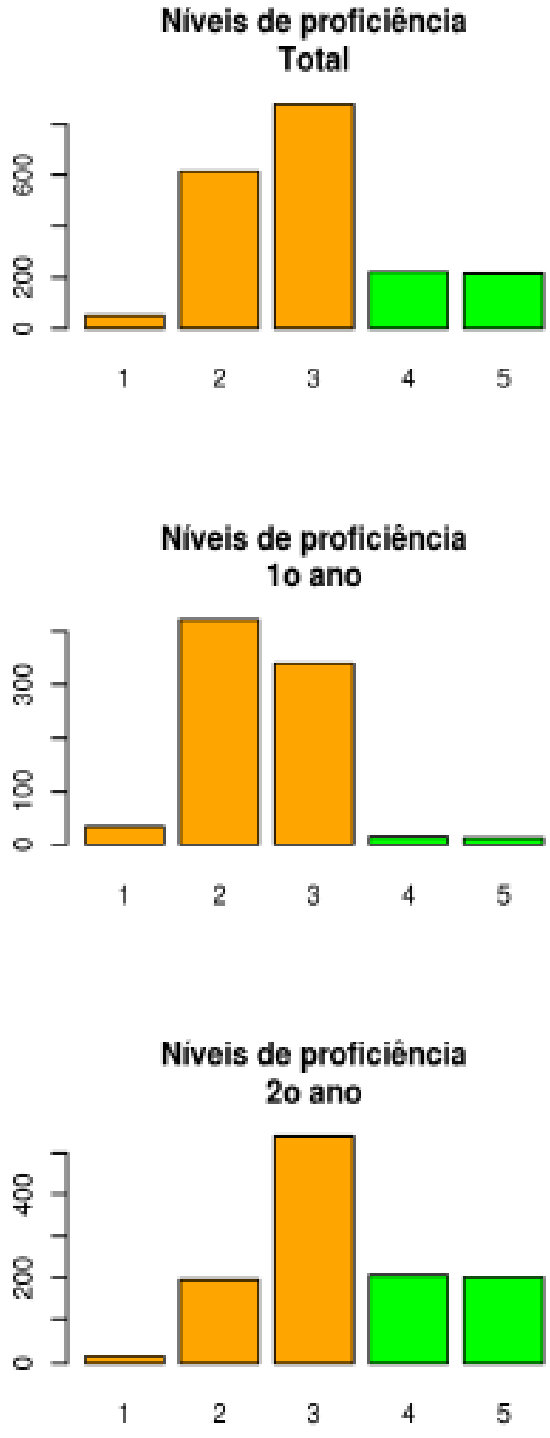
Tabela 17 – Síntese dos resultados da aplicação do TAI da PB – Leitura, por população e subpopulação

	Proficiência	Média de acertos	Número de itens	Duração
TOTAL				
Média	462,72	0,55	17,33	11,52
Desvio padrão	81,93	0,20	2,18	5,53
Dados válidos	1983	1983	1983	1983
1º ANO				
Média	416,83	0,44	16,74	10,88
Desvio padrão	61,41	0,15	1,60	5,75
Dados válidos	823	823	823	823
2º ANO				
Média	495,28	0,63	17,75	11,98
Desvio padrão	79,04	0,19	2,43	5,32
Dados válidos	1160	1160	1160	1160

Fonte: São Paulo (2016d).

Além das proficiências estimadas ao final do teste, o algoritmo alocava o(a) aluno(a) em um dos níveis da escala de proficiência. O sumário dos resultados obtidos para os 1º anos, 2º anos e para todos(as) os(as) respondentes pode ser visto na Figura 31.

Figura 31 – Distribuição dos(as) alunos(as) no TAI da PB – Leitura, por níveis de proficiência da escala e subpopulação



Fonte: São Paulo (2016d).

Importa ressaltar que o resultado fornecido imediatamente ao final do teste não é um atributo exclusivo do TAI, pois o TBC da PB – Leitura, já permitia essa facilidade. De todo modo, considerando que os(as) professores(as) diagnosticam a competência

leitora para propor situações de aprendizagem que possam ser suficientemente desafiadoras, a possibilidade de resultados imediatos é igualmente vantajosa para a rapidez das análises e proposição de intervenções. Além disso, esse resultado pode ser compartilhado com pais, mães ou responsáveis, estudantes e gestores(as) da escola, sendo que o *software* permite adequar a linguagem e as informações em cada situação de uso. Outra possibilidade quanto aos resultados está em permitir os registros longitudinais de informações dos(as) respondentes. Informações sobre resultados de proficiências anteriores podem implicar em novas maneiras de propor o item do início do TAI da PB – Leitura, bem como o monitoramento das ações e intervenções já realizadas. Os aspectos relativos aos possíveis relatórios não puderam ser aprofundados e merecem desdobramentos em pesquisas futuras.

Outro benefício que o TAI possibilita é o de reduzir a quantidade de itens (comprimento do teste). A diminuição na extensão do teste é um efeito vantajoso, uma vez que impede que os(as) alunos(as) se cansem ao responderem um teste longo, o que pode causar o denominado “efeito cansaço”, que intervém de forma especial nos(as) respondentes de menor proficiência.

No TAI da PB – Leitura, verificou-se que o teste foi reduzido para 70,95% dos(as) alunos(as) participantes, vide Tabela 18. Esse percentual corresponde a 58,28% dos(as) alunos(as) dos 2º anos e 88,82% dos(as) alunos(as) dos 1º anos, mostrando que os(as) alunos(as) dos 1º anos foram mais beneficiados(as) com a diminuição do comprimento do teste.

Tabela 18 – Frequência de testes no TAI da PB – Leitura, por comprimento do teste e por subpopulação (1º e 2º anos)

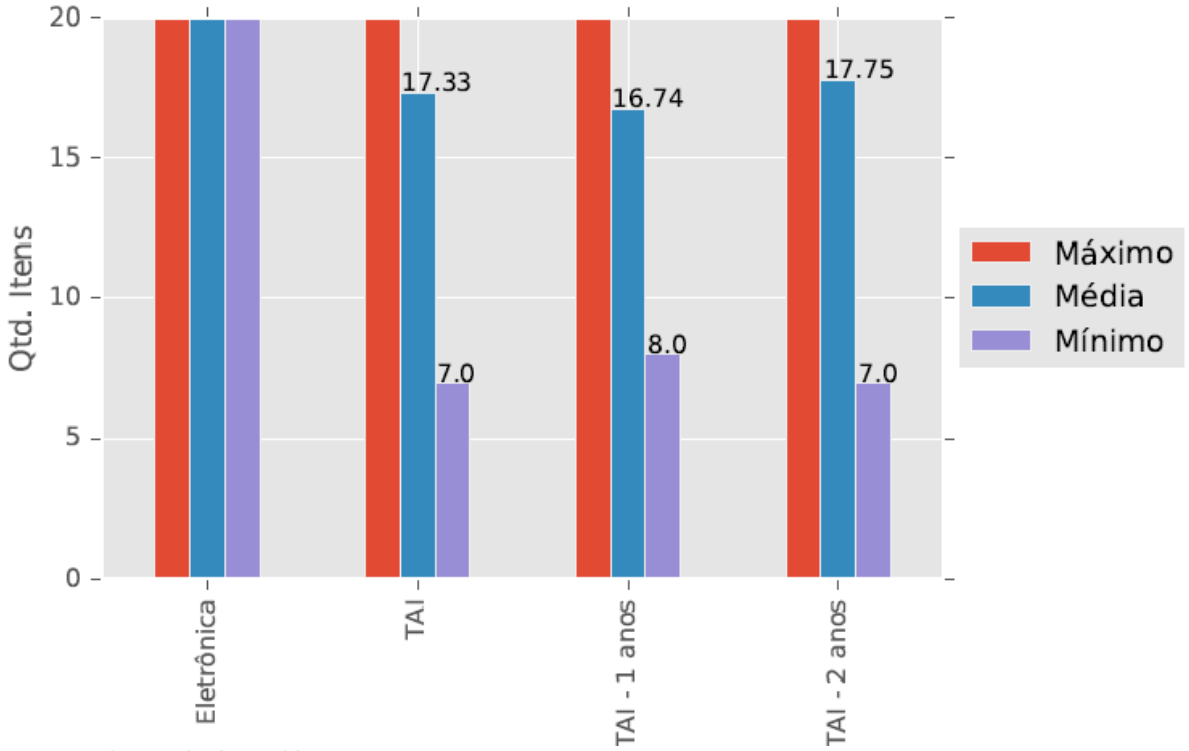
Comprimento do teste	Frequência absoluta 1º ano	Frequência relativa % 1º ano	Frequência absoluta 2º ano	Frequência relativa % 2º ano	Frequência absoluta total	Frequência relativa % total
7	-	-	6	0,52	6	0,30
8	3	0,36	20	1,72	23	1,16
10	7	0,85	4	0,34	11	0,55
15	27	3,28	39	3,36	66	3,33
16	477	57,96	378	32,59	855	43,12
17	131	15,92	100	8,62	231	11,65
18	59	7,17	117	10,09	176	8,88
19	27	3,28	12	1,03	39	1,97
20	92	11,18	484	41,72	576	29,05
Total	823	100,00	1160	100,00	1983	100,00

Fonte: Dados organizados pela autora, com base nos arquivos fornecidos pela SME/SP 2016.

Outro dado que corrobora essa afirmação é apresentado na Figura 32, na qual é mostrada a média de itens respondidos no TBC e no TAI da PB – Leitura. A média de itens respondidos no TAI da PB – Leitura é de 17,33 contra 20 itens para o TBC ou o teste impresso. Essa diminuição representa um percentual de 13,35% da extensão do teste em papel e lápis.

No TAI, verificou-se que os estudantes do 2º ano fizeram provas um pouco maiores, sendo 17,75 itens contra 16,74 do 1º ano.

Figura 32 – Quantidade máxima, mínima e média de itens respondidos, por tipo de aplicação



Fonte: São Paulo (2016b).

Além de testes de menor comprimento, o algoritmo do TAI da PB – Leitura ainda buscou majorar a precisão da medida, dado que as regras que compõem o critério de encerramento do teste são responsáveis por agregar precisão, contribuindo para melhores decisões pedagógicas.

Na Tabela 19, é possível perceber que 70,98% dos respondentes são beneficiados com testes de comprimento menor que 20 itens no TAI da PB – Leitura.

Observa-se também que o teste encerrou com 20 itens para 29,02% do total de respondentes. Esse percentual pode ser decomposto em 7,87% de respondentes que, além de atingirem o limite de 20 itens, também tiveram erro de medida menor que 35

pontos e 21,18% que encerraram o teste somente pela regra do alcance do limite de 20 itens.

Independentemente de terem sido encerrados pela regra do limite de 20 itens, o TAI da PB – Leitura, proporcionou que o(a) respondente recebesse 20 itens mais adequados aos seus domínios, quanto aos parâmetros de dificuldade, contrariamente a uma prova impressa, em que os itens precisam se adequar aos diferentes domínios possíveis na população. As taxas de acerto constituem evidências desse aspecto, conforme será verificado mais adiante.

Tabela 19 – Frequência de testes no TAI da PB – Leitura, por tipo de encerramento e por subpopulação (1º e 2º anos)

Regras de encerramento	Freq. abs. 1º ano	Freq. % 1º ano	Freq. abs. 2º ano	Freq. % 2º ano	Freq. abs. total	Freq. % total
Limite de 20 itens (a)	39	4,74	381	32,84	420	21,18
Erro de medida ≤ 35 pontos (b)	710	86,27	632	54,48	1342	67,68
Inexistência de ponto de corte no intervalo de confiança (c)	15	1,82	41	8,88	56	2,82
(a) + (b)	53	6,44	103	3,53	156	7,87
(b) + (c)	6	0,73	3	0,26	9	0,45
Total	823	100,00	1160	100,00	1983	100,00

Fonte: Dados organizados pela autora, com base nos arquivos fornecidos pela SME/SP 2016.

Isolando os testes que encerraram com 20 itens, a Tabela 20 mostra a variação dos erros de medida, sendo que 27,08% desses testes ainda apresentaram erros em patamares inferiores a 35 pontos.

Tabela 20 – Frequência de testes com encerramento por máximo de 20 itens no TAI da PB – Leitura, por erro de medida e por subpopulação (1º e 2º anos)

Erro de medida (em pontos na escala)	Freq. abs. 1º ano	Freq. % 1º ano	Freq. abs. 2º ano	Freq. % 2º ano	Freq. abs. total	Freq. % total
33,5 --| 35,0	53	57,61	103	21,28	156	27,08
35,0 --| 36,0	7	7,61	6	1,24	13	2,26
36,0 --| 36,5	9	9,78	109	22,52	118	20,49
36,5 --| 37,0	5	5,43	6	1,24	11	1,91
37,0 --| 37,5	1	1,09	1	0,21	2	0,35
38,0 --| 38,5	6	6,52	88	18,18	94	16,32
39,5 --| 40,0	1	1,09	2	0,41	3	0,52
41,0 --| 41,5	6	6,52	79	16,32	85	14,76
44,5 --| 45,0	3	3,26	62	12,81	65	11,28
50,0 --| 50,5	1	1,09	28	5,79	29	5,03
Total	92	100,00	484	100,00	576	100,00

Fonte: Dados organizados pela autora, com base nos arquivos fornecidos pela SME/SP 2016.

Ainda quanto à Tabela 19, percebe-se que a maioria dos testes (75,55%), sendo 58,01% dos testes do 2º ano e 92,71% dos testes do 1º ano, encerrou pela regra do erro de medida.

Excetuando os testes encerrados pela regra de classificação da proficiência, por terem a precisão apoiada por outro aspecto, observou-se a variabilidade para o erro padrão da população submetida ao TAI, em que a média foi de 35,69 e o desvio padrão 2,97 pontos, evidenciando a homogeneidade na precisão das estimativas.

Por sua vez, foram isolados os testes que foram encerrados pela regra de classificação da proficiência, que constituem 3,27% do total de testes, para verificação, na Tabela 21, das dimensões desses erros. Embora seja baixo o percentual de testes que encerraram com patamares de erro inferiores a 35, a precisão foi interpretada nesses testes pelo fato de o intervalo de confiança[53] estar contido em um dos níveis da escala de proficiência, tornando mais preciso o diagnóstico pedagógico.

[53] Conforme já apontado, foi considerado o intervalo de confiança, quando deveria ter sido considerado o intervalo de confiança bayesiano.

Tabela 21 – Frequência de testes encerrados pela regra de classificação no TAI da PB – Leitura, por erro de medida e por subpopulação (1º e 2º ano)

Erro de medida (em pontos na escala)	Freq. abs. 1º ano	Freq. % 1º ano	Freq. abs. 2º ano	Freq. % 2º ano	Freq. abs. total	Freq. % total
34,5 --| 35,0	6	28,57	3	6,82	9	13,85
40,0 --| 45,0	5	23,80	5	11,36	10	15,38
50,0 --| 55,0	7	33,34	10	22,73	17	26,15
60,0 --| 62,0	3	14,28	26	59,09	29	44,62
Total	21	100,00	44	100,00	65	100,00

Fonte: Dados organizados pela autora, com base nos arquivos fornecidos pela SME/SP 2016.

A Tabela 22 demonstra que o encerramento pela regra de classificação garantiu a diminuição do teste para todos os(as) respondentes submetidos a ela, sendo que para 61,53% deles o teste foi reduzido ao número entre 7 e 10 itens.

Tabela 22 – Frequência de testes encerrados pela regra de classificação no TAI da PB – Leitura, por comprimento do teste e por subpopulação (1º e 2º anos)

Comprimento do teste	Freq. abs. 1º ano	Freq. % 1º ano	Freq. abs. 2º ano	Freq. % 2º ano	Freq. abs. total	Freq. % total
7	-	0,00	6	13,63	6	9,23
8	3	14,28	20	45,46	23	35,38
10	7	33,34	4	9,09	11	16,92
15	4	19,05	1	2,27	5	7,69
16	6	28,57	4	9,09	10	15,38
17	1	4,76	9	20,46	10	15,38
Total	21	100,00	44	100,00	65	100,00

Fonte: Dados organizados pela autora, com base nos arquivos fornecidos pela SME/SP 2016.

O TAI da PB – Leitura, apresentou erros de medida de menor variabilidade, conforme simulação, mais baixos que o teste de 20 itens impresso.

Conferir maior precisão às estimativas da proficiência, permitindo que o teste seja mais informativo do ponto de vista psicométrico traz consequências importantes do ponto de vista pedagógico, envolvendo decisões mais assertivas no processo de aprendizagem (STECHER, 2002; STIGGINS, 2008).

Thissen (2014), Steinberg, Thissen e Wainer (2014) e Weiss (2011) destacam que, para os testes impressos, as proficiências estimadas pela TRI variam em relação ao nível de precisão e são mais precisas para examinandos(as) alocados(as) em pontos próximos ao chamado “pico” do teste, em outras palavras, para as proficiências

estimadas na vizinhança dos pontos de maior informação do teste; mas para os TAI, os erros de medição obtidos podem ser mais homogêneos, ainda que os pressupostos da TRI corroborem a variabilidade dos erros de medida.

As consequências do destaque desses autores podem ser vistas na Tabela 23, em que é possível verificar essa homogeneidade, pois dos 1.983 testes realizados, 1.649 (83,15%) apresentaram patamares de erro de medida entre 34 e 37 pontos da escala. Ainda é possível verificar que dos 150 testes com erros de medida de maior dimensão (entre 44 e 62 pontos), 56 (37,33%) foram obtidos com base na regra de classificação da proficiência em um dos níveis da escala, os quais não visaram a precisão da proficiência, mas a precisão da categorização do domínio do(a) estudante.

Tabela 23 – Frequência de testes no TAI da PB – Leitura, por erro de medida e por regra de parada

Erro de medida no TAI da PB – Leitura	Frequência absoluta de testes	Critério de parada (freq. absoluta)				
		Limite de 20 itens (a)	Erro de medida <35 (b)	Intervalo de confiança (c)	(a) +(b)	(b) +(c)
61 |-- 62	29	-	-	29	-	-
52 |-- 53	11	-	-	11	-	-
50 |-- 51	35	29	-	6	-	-
44 |-- 45	75	65	-	10	-	-
41 |-- 42	85	85	-	-	-	-
39 |-- 40	3	3	-	-	-	-
38 |-- 39	94	94	-	-	-	-
37 |-- 38	2	2	-	-	-	-
36 |-- 37	129	129	-	-	-	-
35 |-- 36	13	13	-	-	-	-
34 |-- 35	1424	-	1293	-	122	9
33 |-- 34	83	-	49	-	34	-
Total	1983	420	1342	56	156	9

*n/c (não consta).
Fonte: Dados organizados pela autora, com base nos arquivos fornecidos pela SME/SP 2016.

Outro aspecto apontado como vantajoso nos TAI está na possibilidade de reduzir o tempo de aplicação do teste. Quanto à redução do tempo de aplicação do TAI em relação ao TBC da PB – Leitura, a Tabela 24 exibe o tempo médio de aplicação no teste e no item por escola. Em relação à média do tempo médio do teste, observou-se uma diminuição do TBC para o TAI da PB – Leitura de 3 minutos e 41 segundos. Sugere-se que essa diminuição é assegurada pela diminuição do comprimento dos testes, a julgar pelo número médio de itens nas 1983 aplicações, que ficou no valor

de 17,33 itens (vide Figura 32). Quanto à diminuição no tempo médio por item, de 5 segundos, pode ser justificada na adequação dos itens aos domínios dos(as) alunos(as), resultando em itens menos difíceis para alunos(as) com proficiências baixas e mais difíceis (com textos maiores) para alunos(as) mais proficientes, aspecto que também será evidenciado na taxa de acerto no TAI da PB – Leitura.

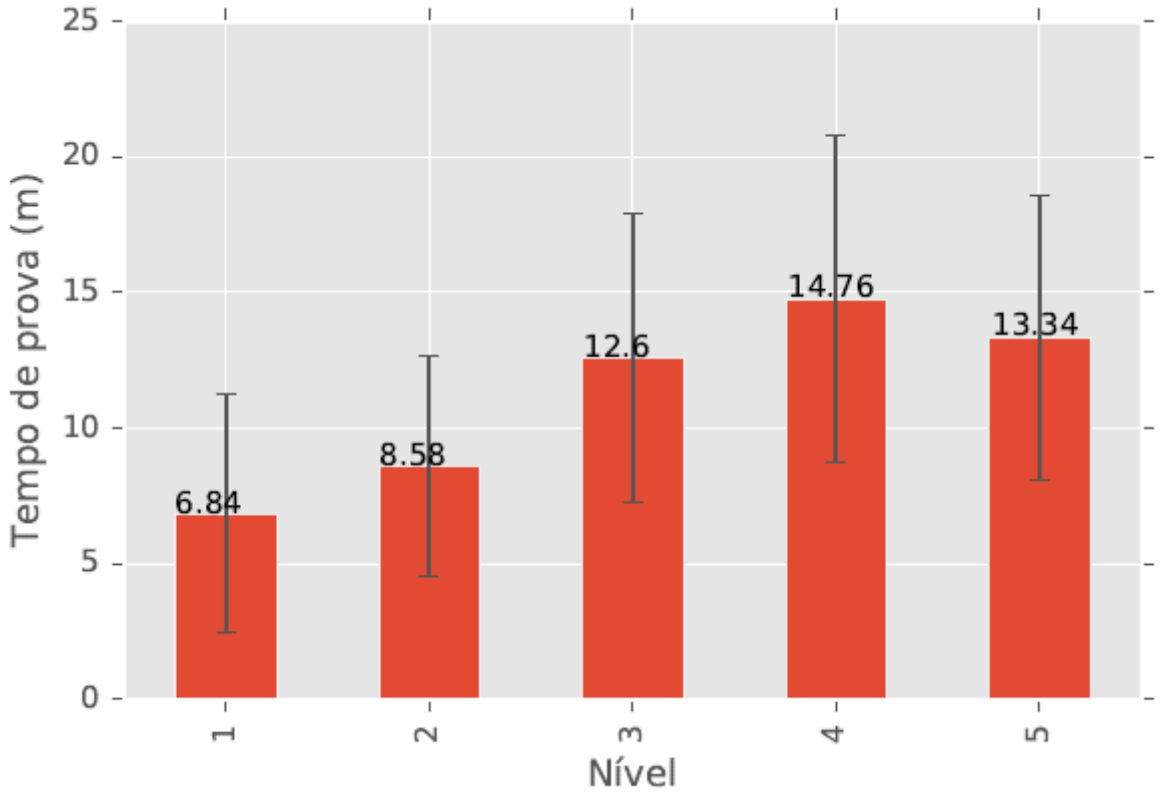
Tabela 24 – Tempo médio no teste e no item para o TBC e para o TAI da PB – Leitura, por escola

Escola	Tempo médio no teste – TBC	Tempo médio por item – TBC	Tempo médio no teste – TAI	Tempo médio por item – TAI
Angola	00:17:05	00:00:48	00:13:24	00:00:42
África do Sul	00:15:13	00:00:43	00:11:24	00:00:37
Moçambique	00:18:00	00:00:51	00:11:14	00:00:37
Libéria	00:13:27	00:00:39	00:11:01	00:00:36
Argélia	00:10:54	00:00:31	00:10:36	00:00:41
Costa do Marfim	00:20:50	00:00:59	00:10:29	00:00:36
Líbia	00:13:34	00:00:38	00:11:32	00:00:38
Cabo Verde	00:12:01	00:00:34	00:11:18	00:00:37
República do Congo	00:13:14	00:00:37	00:11:06	00:00:39
Camarões	00:14:01	00:00:40	00:13:05	00:00:42
Etiópia	00:15:54	00:00:45	00:12:49	00:00:39
Benim	00:15:10	00:00:43	00:10:41	00:00:39
Ruanda	00:16:34	00:00:50	00:11:25	00:00:39
Egito	00:12:28	00:00:35	00:08:25	00:00:31
Marrocos	00:16:11	00:00:46	00:11:10	00:00:37
Média	00:14:58	00:00:43	00:11:17	00:00:38

Fonte: São Paulo (2016b).

Comparando as informações da Figura 12, na subseção 4.3.2, e da Figura 33 a seguir, que exibem as médias de tempo, em minutos, para a conclusão do TBC e do TAI da PB – Leitura, por nível da escala de proficiência, percebe-se que houve diminuição do tempo em todos os níveis da escala, sendo que as diminuições são maiores nos níveis 1 e 5. Assim como no TBC, os(as) alunos(as) de menor proficiência continuam a concluir o teste em menor tempo. Para o TBC, o menor tempo pode ter ocorrido devido à inexistência de competência leitora, acarretando a imediata desistência em ler o item e responder-lhe. Para o TAI, o menor tempo, além de ser relativo à desistência em ler enunciados e responder a itens que exigem maior competência leitora, está associado à menor quantidade de itens com essas características selecionados para o(a) respondente, pois o algoritmo de seleção tenderia a trazer itens mais fáceis diante das sucessivas respostas erradas.

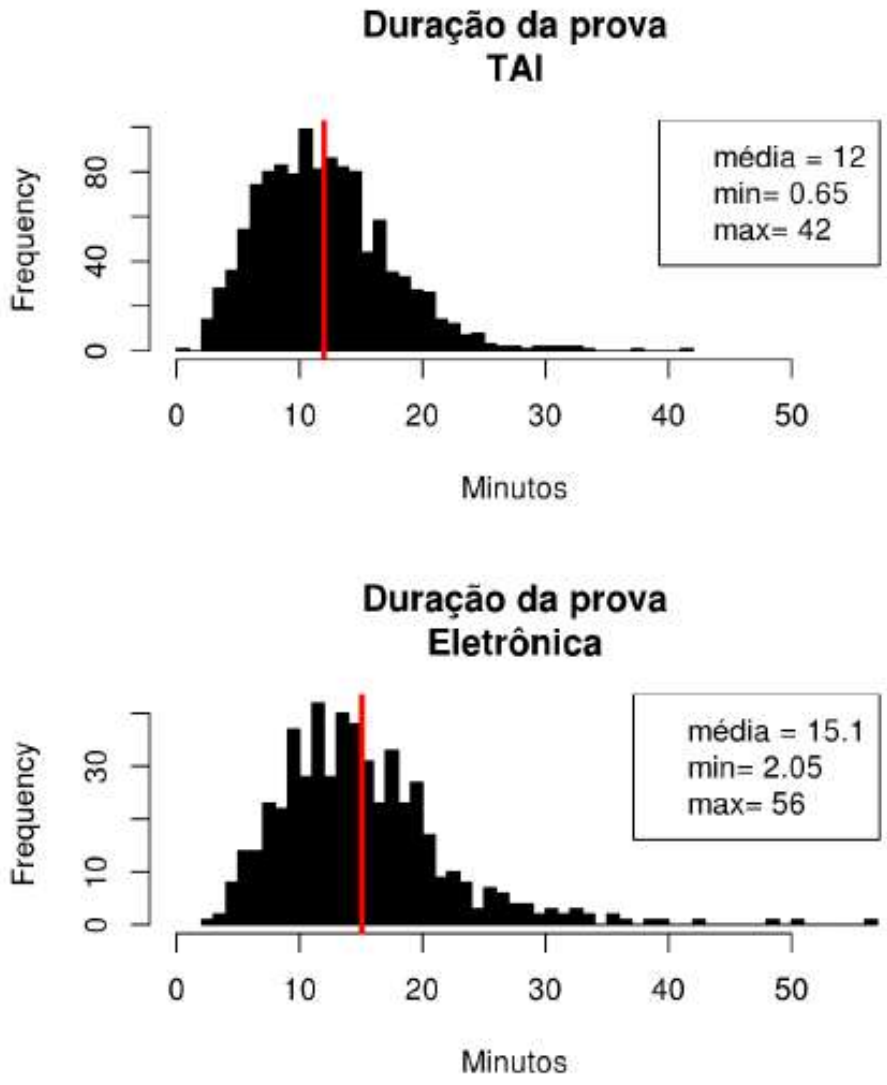
Figura 33 – Média de tempo, em minutos, para conclusão do TAI da PB – Leitura, por nível da escala



Fonte: São Paulo (2016b).

Considerados o tempo médio de duração para a aplicação do teste impresso (vide Tabela 7) de 38 minutos e o tempo médio de duração do teste no TAI da PB – Leitura, (vide Figura 34), de 12 minutos, revelou-se uma diminuição de 26 minutos que corresponde a 68,42%. Tomados os valores de tempo médio de duração no TBC e no TAI, constata-se uma diminuição de cerca de 3 minutos.

Figura 34 – Média de duração do teste, em minutos, por tipo de administração



Fonte: São Paulo (2016d).

A diminuição no tempo do teste permite que a aplicação do instrumento ocupe menor tempo do período de aula e também que mais testes possam ser aplicados durante o período letivo, considerando o uso de caráter formativo.

A personalização do teste é outra vantagem que um TAI promete. No teste adaptativo, conforme o domínio do(a) respondente, ele(a) recebe um teste ajustado aos seus conhecimentos. Esse ajuste pode significar maior taxa de êxito no teste, devido à seleção de itens com parâmetros “b” próximos ao valor das proficiências estimadas, ou seja, em nível de dificuldade mais próximo ao domínio do(a) respondente. A despeito de propiciar testes mais precisos, essa característica certamente tem consequências na motivação para a realização do teste.

Uma forma de observar esse ajuste está em comparar a taxa de acertos no TAI da PB – Leitura e do teste em papel. No teste da PB – Leitura em papel e lápis, a quantidade de acertos que aloca o participante em determinado nível da escala foi apresentada pelo Inep e mostrada na Tabela 13, da subseção 4.4.2. Nessa tabela, para o(a) aluno(a) ser alocado no nível 1 da escala, ele deveria apresentar até três acertos nas 20 questões do teste, representando uma taxa de 15% de acertos. Comparando essa taxa com os dados da Tabela 25, verifica-se a existência de 22 alunos(as) com taxas acerto maiores (esses valores aparecem em negrito) e ainda alocados no nível 1, significando que seus testes estavam mais ajustados aos seus domínios, tendo como consequência maior taxa de acerto. Esses 22 alunos(as) representam um percentual de 42,3% dos(as) alunos(as) do nível 1 com maiores taxas de acerto no TAI da PB – Leitura.

Consequentemente, para o(a) aluno(a) ser alocado(a) no nível 2 da PB – Leitura em papel e lápis, deveria acertar entre 20 e 35% das 20 questões do teste. Na Tabela 26, verifica-se que, 411 respondentes tiveram suas taxas de acerto maiores, evidenciando que o TAI apresentou testes mais ajustados aos seus conhecimentos e, por sua vez, melhores taxas de acerto para 66,72% dos(as) respondentes do nível 2.

Tabela 25 – Distribuição de testes do TAI da PB – Leitura, por taxa de acerto e nível de desempenho

Taxa de acerto (%)	Número de testes					
	Nível 1	Nível 2	Nível 3	Nível 4	Nível 5	Total
95 --| 100	-	-	-	-	56	56
90 --| 95	-	-	-	-	72	72
85 --| 90	-	-	-	-	80	80
80 --| 85	-	-	-	**91**	6	97
75 --| 80	-	-	**1**	**122**	-	123
70 --| 75	-	-	**92**	**8**	-	100
65 --| 70	-	-	**139**	-	-	139
60 --| 65	-	-	**87**	-	-	87
55 --| 60	-	-	182	-	-	182
50 --| 55	-	**47**	162	-	-	209
45 --| 50	-	**107**	156	-	-	263
40 --| 45	-	**135**	57	-	-	192
35 --| 40	-	**122**	4	-	-	126
30 --| 35	-	74	-	-	-	74
25 --| 30	-	95	-	-	-	95
20 --| 25	**4**	34	-	-	-	38
15 --| 20	**17**	2	-	-	-	19
10 --| 15	10	-	-	-	-	10
05 --| 10	13	-	-	-	-	13
00 --| 05	8	-	-	-	-	8
Total	52	616	880	221	214	1983

*Em negrito os testes cuja taxa de acerto no TAI da PB – Leitura foi superior ao teste impresso.
Fonte: Dados organizados pela autora, com base nos arquivos fornecidos pela SME/SP 2016.

Para alunos(as) alocados(as) no nível 3 da PB – Leitura em papel e lápis, os acertos deveriam estar entre 40 e 60% das 20 questões do teste. Na Tabela 25, revela-se que 319 crianças tiveram suas taxas de acerto maiores, evidenciando que o TAI apresentou maiores taxas de acerto para 36,25% dos(as) examinandos(as) desse nível.

As taxas de acerto consideradas no teste impresso ficaram entre 65 e 75% do teste para o nível 4 e entre 80 e 100% do teste para o nível 5. Desse modo, evidenciou-se que todos os(as) 221 alunos(as) do nível 4 submetidos ao TAI da PB – Leitura tiveram taxas de acerto maiores que o teste impresso e, no nível 5, houve equivalência das taxas de acerto nas duas formas de administração para todos os testes.

Em suma, para 972 respondentes (49% do total) as taxas de acerto foram melhores no TAI e para os demais as taxas de acerto foram idênticas ao teste impresso.

A Tabela 26 também revela aspectos sobre as taxas de acerto, acrescentando informações sobre o comprimento do teste por níveis de proficiência.

Tabela 26 – Distribuição dos testes do TAI da PB – Leitura, por comprimento do teste e níveis de proficiência

Comprimento do teste	Número de testes					
	Nível 1	Nível 2	Nível 3	Nível 4	Nível 5	Total
7	-	-	-	-	6	6
8	-	-	-	-	23	23
10	11	-	-	-	-	11
15	5	1	60	-	-	66
16	1	381	473	-	-	855
17	4	131	90	-	6	231
18	-	53	123	-	-	176
19	-	28	11	-	-	39
20	31	22	123	221	179	576
Total	52	616	880	221	214	1983

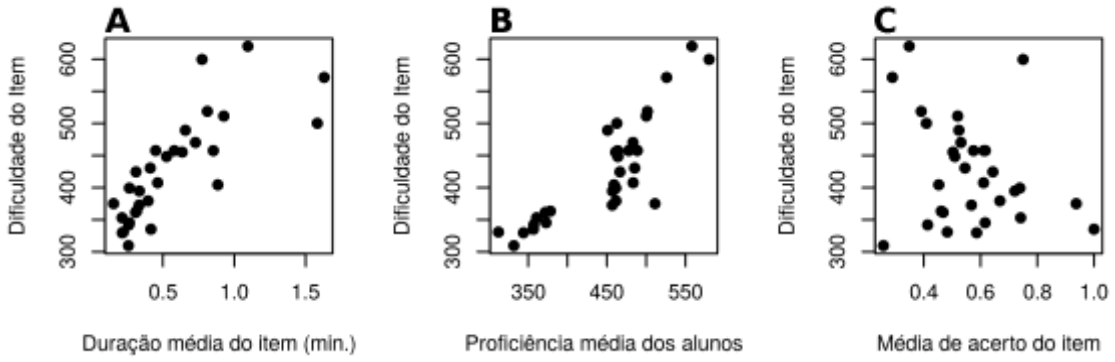
Fonte: Dados organizados pela autora, com base nos arquivos fornecidos pela SME/SP 2016.

Por meio dela, é possível notar que os(as) respondentes submetidos(as) ao TAI da PB – Leitura que tiveram testes de menor comprimento (7 ou 8 itens) são alocados(as) no nível 5, sendo que foi observado, nos microdados, que esses respondentes tiveram 100% de taxa de acerto no teste. Também é importante destacar que os(as) respondentes com 0% de taxa de acerto no teste (8 estudantes) responderam a 10 itens e foram alocados(as) no nível 1. Esse aspecto mostra que o TAI solucionou o problema em relação à aplicação do teste impresso da PB – Leitura, já exemplificado, no qual um(a) aluno(a) com proficiência do nível 1 tinha que continuar a responder aos itens de dificuldade maior, caracterizando uma situação desconfortável, enquanto um(a) aluno(a) com proficiência no nível 5, mesmo que desmotivado, tinha que responder a itens muito fáceis, do início da prova.

Essa informação pode ser corroborada na Figura 35. A dispersão ilustrada em 35B confirma que no TAI da PB – Leitura os itens mais difíceis foram administrados aos(às) alunos(as) com maior proficiência. Em contrapartida, não se observa relação linear entre dificuldade dos itens e a média de acerto (35C), tradicional nos testes impressos, confirmando uma marca dos testes adaptativos em direcionar a apresentação do item ao público que pode responder-lhe.

Também se demonstra na Figura 35A que os itens mais difíceis exigiram maior tempo de resposta.

Figura 35 – Dispersão entre a dificuldade (parâmetro b) dos itens do TAI da PB – Leitura e três aspectos de sua aplicação: A) o tempo médio para resolução do item; B) a proficiência média dos alunos que receberam o item; C) a média de acerto no item



Fonte: São Paulo (2016d).

Em grande medida, a qualidade de um TAI está na qualidade do BI que dispõe. BI muito pequenos e com itens pouco discriminativos podem resultar em testes que não alcançam os objetivos estabelecidos nas etapas do algoritmo, maior precisão e menor comprimento do teste e sua adequação ao domínio do(a) respondente. Além disso, com o BI restrito, os itens são repetidamente administrados aos(às) diferentes respondentes, causando uma exposição que pode levar futuros(as) respondentes a acertarem o item sem que apresente o domínio correspondente ao acerto. Esses problemas são tratados como a melhoria da segurança e da taxa de exposição dos itens do teste.

O BI limitado a 39 itens impossibilitou a inserção de regras de controle das taxas de exposição no critério de seleção de itens, que são importantes por assegurar que os testes contínuos – aqueles aplicados com maior frequência no ano letivo, por se constituírem ferramenta de diagnóstico do domínio de habilidades –, não sofram o viés de acerto por conhecimento prévio dos itens, fazendo com que os(as) respondentes recebam estimativas inflacionadas da proficiência, conforme alertam Barrada et al. (2009b, 2011). Esse desafio deverá ser enfrentado em estudos futuros.

Embora os prejuízos com a sobreposição de itens tenham que ser superados, na aplicação do TAI da PB – Leitura, as consequências desse quesito não foram tão ruins, uma vez que todos os(as) respondentes de uma escola realizaram o teste no mesmo dia, sendo que não havia contato entre respondentes que realizaram o TAI e aqueles que ainda não o tinham respondido. Além disso, os estudantes envolvidos

apresentavam idade entre 6 e 7 anos de idade e ainda não encaram os testes como os estudantes mais velhos, de modo que ainda ignoram a prática de “passar cola”.

Quanto à utilização dos itens do BI do TAI da PB – Leitura, mesmo não havendo diferenças nos parâmetros “a” dos itens, foi possível constatar que 10 itens, a maioria no extremo superior da escala, não foram utilizados, caracterizando uma utilização de 74,35% dos itens do banco (Figura 38).

A explicação para isso pode se assentar na constatação de que 16,35% dos(as) respondentes com maior proficiência tiveram seus testes encerrados pela regra de classificação, baseada no intervalo de confiança, aspecto que pode ser observado na Tabela 27 que traz os testes por tipo de encerramento e nível de proficiência.

Tabela 27 – Distribuição dos testes do TAI da PB – Leitura, por comprimento do teste e níveis de proficiência

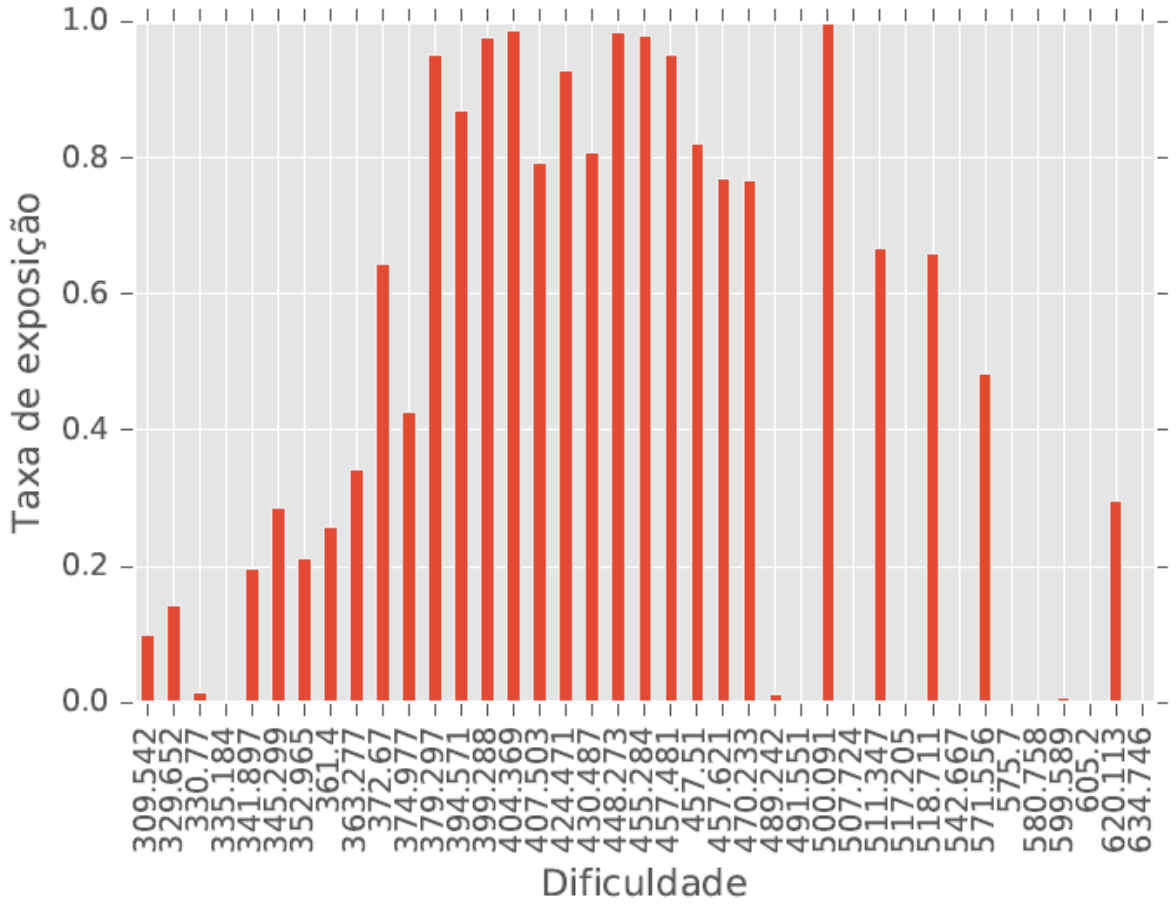
Regras de encerramento	Número de testes por nível de proficiência					
	Nível 1	**Nível 2**	**Nível 3**	**Nível 4**	**Nível 5**	**Total**
Limite de 20 itens (a)	17	2	1	221	179	420
Erro de medida ≤ 35 pontos (b)	-	585	757	-	0	1342
Inexistência de ponto de corte no intervalo de confiança (c)	21	-	-	-	35	56
(a) + (b)	14	20	122	-	-	156
(b) + (c)	-	9	-	-	-	9
Total	52	616	880	221	214	1983

Fonte: Dados organizados pela autora, com base nos arquivos fornecidos pela SME/SP 2016.

Com base nas observações de Veldkamp e van der Linden (2008), nota-se que, embora essa utilização seja baixa, ainda é maior do que os 10% ou mais dos itens em geral usados em BI de testes adaptativos parametrizados pela TRI de três parâmetros, na ausência do controle de exposição.

O item com valor b = 500,091 foi o primeiro item apresentado em todos os TAI, gerando uma taxa de 100% de exposição, vide Figura 36. Outros 10 itens tiveram taxas altas, acima de 80%, de exposição. Percebe-se ainda que a grande maioria dos itens não utilizados se encontra na porção superior da escala de dificuldades dos itens.

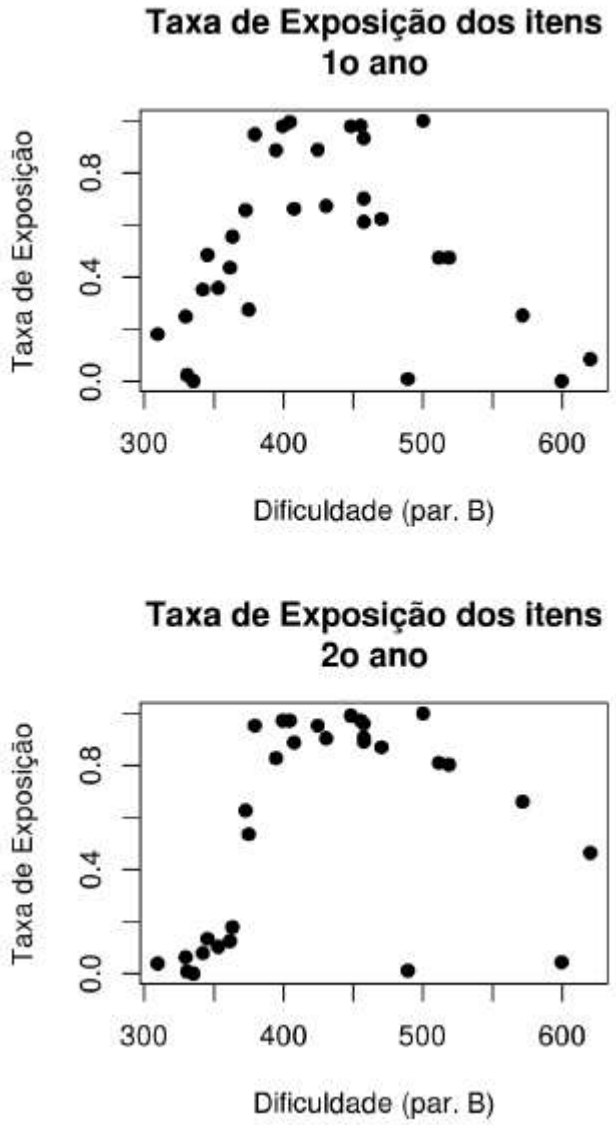
Figura 36 – Taxa de exposição dos itens utilizados no TAI da PB – Leitura



Fonte: São Paulo (2016b).

Também em relação à taxa de exposição dos itens do TAI da PB – Leitura, a Figura 37 busca relacionar a taxa de exposição e a dificuldade do item, por subpopulação de respondente. Os resultados demonstram que, além de uma taxa de exposição maior para os itens de dificuldade média, no 2º ano, houve um uso maior dos itens mais difíceis e um uso menor dos itens mais fáceis.

Figura 37 – Relação entre taxa de exposição e dificuldade dos itens no TAI da PB – Leitura



Fonte: São Paulo (2016d).

Outro aspecto que deve ser considerado quanto ao BI refere-se à terceira regra do critério de encerramento, baseada na classificação. Essa regra requer que maior quantidade de itens tenha o parâmetro de dificuldade nos limiares dos pontos de corte da escala, aspecto que também não pode ser analisado para o TAI da PB – Leitura.

Apesar das aplicações nas 15 unidades terem sido bem-sucedidas, é necessário apontar os desafios técnicos encontrados na aplicação do TAI da PB – Leitura.

O primeiro, referente à base de alunos(as), consistiu na não correspondência entre o cadastro dos(as) alunos(as) no Sistema Escola On-Line (EOL) e as reais situações nas escolas, desencadeando os seguintes problemas:

a) mesmos(as) alunos(as) cadastrados(as) em turmas diferentes na mesma escola;
b) mesmos(as) alunos(as) cadastrados(as) em escolas diferentes;
c) mesmos(as) alunos(as) cadastrados(as) em anos diferentes;
d) alunos(as) não cadastrados.

Nesses casos, a situação correta dos(as) alunos(as) era inferida por meio dos dados disponíveis, por exemplo, presença de número na chamada ou data da matrícula mais recente.

O segundo, referente à infraestrutura das escolas necessária para o TAI. A SME/SP tinha 55 *tablets* para aplicação do TAI e para que pudessem acessar o *Wi-Fi* das escolas, seus endereços *Media Access Control*[54] (MAC) tiveram que ser cadastrados individualmente no *firewall* das 15 escolas. Ademais, foi descoberto um limite máximo no número de conexões por *access point*[55] (AP) e, ao chegar a aproximadamente 30 conexões sem fio, o AP falhava em autenticar mais aparelhos. Portanto, foi necessário, por recomendação do setor de gerenciamento de tecnologias da SME/SP, que a aplicação fosse feita em locais com cobertura de, no mínimo, dois AP ou separando os *tablets* em duas salas de aula sob cobertura de AP diferentes.

Em terceiro, não houve possibilidade de cadastro dos 55 endereços MAC dos *tablets* no *firewall* de todas as escolas, de forma que, em geral, 10 *tablets* não podiam ser utilizados, restrição que ocasionou o uso de uma estratégia, na qual as aplicações eram realizadas para uma turma por vez, dividindo essa turma entre as salas onde os *tablets* estavam instalados. A escolha das salas que seriam usadas para instalação era feita no dia da aplicação, incluindo a sala de informática, a sala de leitura ou uma sala de aula. Os(as) alunos(as) eram levados(as) às salas escolhidas no momento da aplicação e levados(as) de volta às salas de aula assincronamente, conforme terminavam a prova TAI.

O quarto desafio se referiu ao volume dos fones de ouvido. Os fones originais dos *tablets* eram dispositivos auriculares denominados *in-ear* ou “auriculares”, que não atendiam às necessidades higiênicas e de inclusão impostas ao projeto, por isso, foi

---

[54] Endereço físico associado à interface de comunicação, que conecta um dispositivo à rede.
[55] Dispositivo de rede usado para estender a cobertura de redes de Internet. O aparelho funciona conectado via cabo a um roteador – ou um switch – e distribui sinal Wi-Fi na outra ponta.

realizada a compra de *headsets* (Figura 38), ou seja, fones de ouvido com arco. Os fones de ouvido adquiridos possuíam impedância alta (32 Ohm) se comparados com os fones de ouvido auriculares originais do Samsung Galaxy Tab 10.1 (16 Ohm) – ver Figura 38 –, tornando necessário o ajuste do volume para permitir audição satisfatória da voz que narrava o texto dos itens, especialmente nos ambientes pouco silenciosos das escolas. O quinto desafio se referiu às baterias dos *tablets*, visto que o uso constante do *Wi-Fi* e tela ligada, na aplicação do TAI, incorreu no esgotamento das baterias ao fim do dia de aplicação de provas. Durante os intervalos entre aplicações e no translado de uma escola para outra era necessário desligar as telas dos aparelhos e, à noite, era necessário carregar a bateria de todos os *tablets*, período que levava de 4 horas e 30 minutos a 6 horas. Nos dias em que a aplicação envolvia duas ou três escolas, era necessário substituir alguns tablets por aparelhos das próprias unidades escolares.

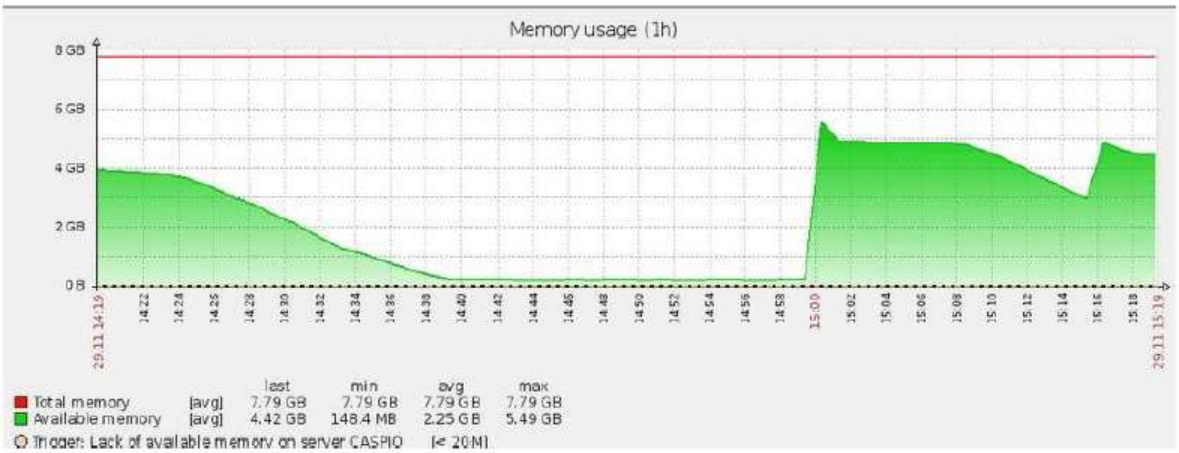
Figura 38 – Imagem do tablet Samsung Galaxy Tab 10.1 P7510 e do headset Multilaser PH002, utilizados nas aplicações do TBC e do TAI da PB – Leitura

Fonte: São Paulo (2016b).

Por fim, o sexto e último desafio estava nas quedas de sistema durante a aplicação do TBC e do TAI. O consumo de memória RAM aumentou para 5 GB na aplicação do TAI da PB – Leitura para uma turma, o que induziu ao aumento da memória RAM da máquina virtual para 8 GB. As Figuras seguintes mostram, respectivamente: o consumo de memória RAM do servidor no fim de uma turma e início de outra, em que

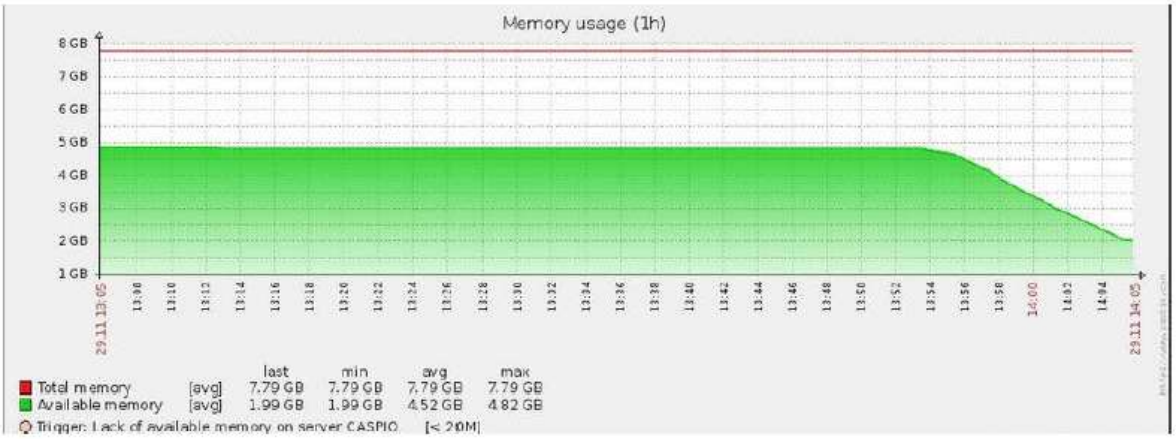
a memória utilizada cai e volta a crescer (Figura 39); o consumo constante de 5 GB de memória RAM durante a aplicação para uma turma (Figura 40); e o uso de memória RAM durante todo o período de aplicação das provas TBC e TAI da PB – Leitura (Figura 41).

Figura 39 – Uso de memória RAM do servidor durante a aplicação do TAI da PB – Leitura para 27 alunos(as): fim de uma turma e início de outra



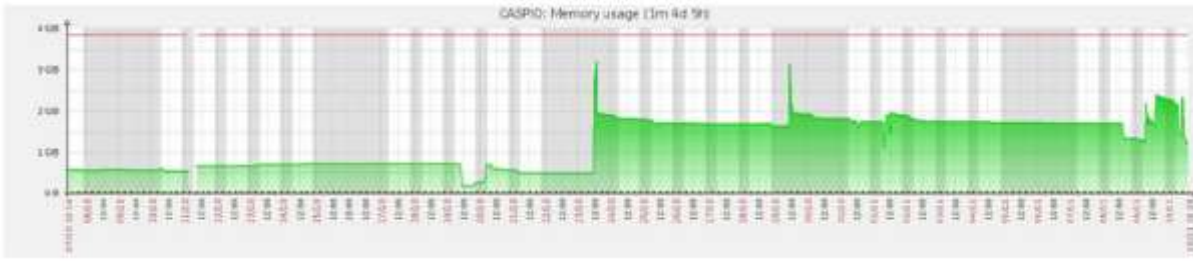
Fonte: São Paulo (2016b).

Figura 40 – Uso da memória RAM do servidor durante a aplicação do TAI da PB – Leitura para 27 alunos(as)



Fonte: São Paulo (2016b).

Figura 41 – Uso de memória RAM do servidor durante as aplicações do TBC e do TAI da PB – Leitura: período de outubro e novembro/2016



Fonte: São Paulo (2016b).

É possível observar que o uso de memória é menor no período de aplicação do TBC da PB – Leitura em relação à aplicação do TAI. O sistema também foi alvo de quedas durante as primeiras aplicações do TAI da PB – Leitura. A causa secundária das quedas foi diagnosticada como falta de memória RAM. Como solução, o sistema de monitoramento de servidores foi configurado para liberar a memória ociosa caso um limiar fosse atingido. O diagnóstico do sistema em busca da causa primária das quedas foi inviabilizado durante as aplicações de provas eletrônicas.

5 CONSIDERAÇÕES FINAIS E PERSPECTIVAS FUTURAS

Os avanços tecnológicos mais recentes têm proporcionado novos horizontes para o campo da educação, mas sua incorporação nas práticas avaliativas ainda é pouco observada no Brasil. Os testes adaptativos informatizados podem representar essa incorporação e, embora sejam mais disseminados em países como Estados Unidos e Espanha, experiências têm sido acanhadamente implementadas no território brasileiro.

O estudo realizou uma aplicação experimental do TAI PB – Leitura, uma versão adaptativa do teste em papel da PB – Leitura, utilizada para diagnóstico da proficiência em leitura. A escolha desse teste se baseou na sua ampla utilização por professores(as) que trabalham, prioritariamente, com a alfabetização e no fato de possuir todas as etapas que configuram um processo de avaliação.

O objetivo fundamental foi verificar se uma versão adaptativa da PB – Leitura poderia apoiar professores(as) e gestores(as) na avaliação da proficiência em leitura de alunos(as) dos anos iniciais, igualmente conseguindo: suprir lacunas inerentes aos processos avaliativos na alfabetização e no letramento inicial, encontrados no teste impresso da PB – Leitura, especialmente, quanto à administração mais rápida, fidedigna e adaptada aos conhecimentos dos(as) respondentes; possibilitar espaço de reflexão para professores(as) e gestores(as) sobre o processo de avaliação da proficiência leitora, especialmente no que se refere aos processos de medida educacional. Procurou-se evitar a todo custo a ideia de que o TAI seria um artefato que, por si só, superaria os desafios de avaliar. Ao contrário, o propósito principal foi garantir que esse TAI se apresentasse como uma ferramenta de apoio ao trabalho docente, sobretudo, em face das tarefas de avaliação, tendo os professores como sujeitos.

Como ponto de partida, perscrutou-se as vantagens que os TAI apresentam mediante revisão da literatura, que também deu destaque às desvantagens dos TAI, como a exigência de amplos bancos de itens, custos altos e a necessidade de estudos que promovam aplicações experimentais do TAI, tendo em vista que grande parte das experiências é decorrentes de simulações.

Os resultados apontaram que foi possível desenvolver a ferramenta TAI da PB – Leitura reforçando as indicações da literatura sobre a necessidade de articulação de um amplo leque de profissionais com diferentes *expertises*, fruto ainda de parcerias e fomentos importantes (Gepave, Secretaria Municipal de Educação de São Paulo e Inep), inclusive pela demanda de altos investimentos.

Foi perceptível a supressão de lacunas dos processos avaliativos na alfabetização e no letramento inicial no processo de construção do artefato. De um lado, o processo de formação possibilitou reflexões e críticas de professores(as) e gestores(as) das 15 escolas, que denotaram preocupações com cada aspecto do processo de medida educacional envolvido no dispositivo TAI da PB – Leitura, em especial, que ele servia para apoiar os(as) educadores(as) na prática avaliativa, sem substituí-los. De outro lado, essa formação confirmou quão necessária é a participação dos(as) professores(as) alfabetizadores(as) nas diferentes etapas de concepção do teste, especialmente nas de interpretação de resultados e sugestões pedagógicas, pois desconheciam aspectos do processo de medida educacional, indicando a ausência de uma política de formação que incorpore a discussão sobre medida educacional como suporte para a avaliação formativa.

Quanto às lacunas do teste impresso da PB – Leitura, foi desenvolvido um TAI da PB – Leitura que gerou testes consistentes, em relação ao arcabouço conceitual da TRI e a metodologia adotada pelo Inep para a elaboração das provas, e cujos resultados corroboraram os parâmetros dos itens produzidos pelo Inep.

Com o emprego do TAI da PB – Leitura, as proficiências estimadas para alunos(as) do 1º e do 2º ano se revelaram consistentes, diante da diferença observada nas médias obtidas, o que mostra que ele pode ser usado inclusive em diferentes populações dos anos iniciais, podendo incluir também o 3º ano, desde que seja ampliado o banco de itens.

Para sua efetiva operacionalização, o TAI dispunha de um algoritmo escrito na linguagem de programação R, com base nos pacotes: catR, PP e irtoys, englobando três componentes: estimação de proficiência, realizada por EAP; seleção do próximo item do teste, por máxima informação de Fisher; e critério de finalização, o qual contou com as seguintes regras:

- limite do número de itens do teste (mínimo de 7 e máximo de 20 itens);

- limite permitido de incerteza (Erro Padrão menor ou igual a 35 pontos);
- intervalo da proficiência (de 85% de confiança) ser incluído no nível de proficiência da escala da PB – Leitura.

A regra do intervalo da proficiência constituiu uma modificação no critério de finalização mais utilizado, pois incorporou ao critério uma regra que é normalmente utilizada em avaliações para fins de certificação ou que procuram categorizar sujeitos em duas ou mais categorias. Essa regra, identificada por regra de classificação da proficiência, se preocupa em alocar o intervalo de confiança da proficiência estimada em um dos cinco níveis da escala de proficiência da PB – Leitura. Ao proporcionar a alocação do intervalo em um nível da escala, diferencia os domínios de leitura dos(as) alunos(as) e, por conseguinte, apoia o processo de avaliação e de intervenção pedagógica. É importante registrar que não foram construídos intervalos de confiança bayesianos, mais adequados aos processos de estimação implementados, ainda que esse aspecto não tenha invalidado o processo de construção do TAI, objeto desta tese.

Embora apenas 3,3% dos(as) respondentes tenham se beneficiado com essa regra, o uso do algoritmo do TAI da PB – Leitura supriu lacunas do teste impresso e ainda foi possível proporcionar para 1.983 alunos(as), 80 professores(as) e 30 gestores(as) de escolas:

a) aplicação informatizada, contribuindo para o uso da tecnologia na avaliação educacional, a padronização da leitura dos itens na aplicação do teste e eliminação de trabalho de pontuação e tabulação dos acertos das provas impressas;
b) pontuação e nível da escala imediatamente após o término da aplicação, eliminando a necessidade de análise dos níveis de proficiência correspondentes aos acertos;
c) flexibilização de horário e dia da administração do teste, podendo ocorrer para um indivíduo ou um grupo e em diferentes momentos do ano, pois não necessita de providências quanto à reprodução dos cadernos de prova e o tempo que ocupa da aula é relativamente menor (12 minutos em média);

d) testes adaptados quanto ao nível de dificuldade, comprovado pelos 49% de respondentes com as taxas de acerto melhores no TAI da PB – Leitura do que no teste impresso e 51% taxas idênticas ao teste convencional, também se verificou a resolução do problema da prova impressa, no qual, por um lado, respondentes do nível 1 tinham que responder a muitos e difíceis itens e, por outro, respondentes do nível 5 tinham que responder a itens fáceis;
e) testes com número reduzido de itens, uma vez que havia menos de 20 itens no teste para 71% dos participantes e o número médio de itens administrado foi de 17,3, sendo que a regra de encerramento pela regra de classificação, agregada à etapa de finalização do teste, garantiu redução considerável (7 a 10 itens) para 2% dos respondentes, sem comprometer a identificação do domínio e a precisão da intervenção pedagógica;
f) testes mais precisos, pois foi constatado que os erros de medida apresentados no TAI da PB – Leitura foram previstos para configurarem dimensões menores que o teste com 20 itens, por meio de simulações, e apresentarem menor variabilidade. Além do mais, o TAI possibilitou a precisão de relevância pedagógica com os testes finalizados pela regra de classificação, mesmo que para poucos respondentes (3,3%). Para esses casos, mesmo o erro tendo maior dimensão, todo o intervalo da proficiência de alunos(as) foi incluído em um nível da escala;
g) reduzir o tempo de aplicação, houve uma diminuição do tempo médio de duração da aplicação da prova impressa para o TAI da PB – Leitura (de 26 minutos), que correspondeu a 68,4% no tempo médio;
h) reflexões os aportes teórico-metodológicos da PB – Leitura, dado que o processo de formação para os(as) educadores(as) tratou de aspectos que subsidiaram a prática avaliativa, especificamente, quanto aos limites e às possibilidades da medida poder apoiar um julgamento referente à alfabetização e ao letramento inicial; e
i) adesão de professores e alunos na utilização do TAI da PB – Leitura evidenciou que esse instrumento tem potencial para transformar momentos de avaliação em oportunidades de trabalho escolar que considerados como interessante e efetivamente motivador, quer pelos ganhos apontados para os docentes, quer pelos ajustes das tarefas avaliativas ao nível de conhecimento dos(as) alunos(as).

Embora as aplicações tenham sido realizadas satisfatoriamente, vale lembrar que diversos obstáculos de ordem técnica na aplicação do TAI da PB – Leitura tiveram que ser contornados, quais sejam: inconsistências nos cadastros dos alunos no Sistema da SME/SP, infraestrutura de acesso ao *Wi-Fi* das escolas, restrição no cadastro dos *tablets* para acesso à rede de conexão da escola, restrição do volume dos fones de ouvido, que não eram os originais dos *tablets,* e quedas do sistema.

Sendo assim, o TAI da PB – Leitura alcançou os objetivos delineados e apontou caminhos promissores para a avaliação educacional, especialmente no que tange a um dos maiores desafios da escolarização de amplos contingentes: a avaliação da proficiência em leitura na perspectiva do sucesso de todos os(as) alunos(as) quanto a essa competência.

Como perspectivas futuras, alguns aspectos merecem ser destacados. Quanto à disponibilização imediata dos resultados ao término da prova, cabe refletir com professores(as) e gestores(as) a incorporação de diferentes níveis de relatórios, considerando perfis diferentes de usuários, como professores(as), alunos(as), gestores(as) da escola, dos órgãos centrais e os pais, mães ou responsáveis. As informações desses relatórios facultam, além da avaliação propriamente dita, diagnósticos diretamente relacionados aos processos de ensino e aprendizagem em sala de aula, essenciais para (re)direcionar o trabalho do(a) professor(a) ou para indicar intervenção individualizada a determinados(as) alunos(as).

Adicionalmente, os resultados das aplicações devem constituir um repositório para armazenar informações longitudinais de modo que permitam um acompanhamento *pari passu* das aprendizagens e, inclusive, possibilitar seu uso em novas aplicações definindo critérios para iniciar o teste, pautando a seleção do item inicial com referência na última proficiência estimada.

O TAI foi uma versão do teste impresso para que a estimação pudesse apoiar-se em uma escala de proficiência interpretada, algo seminal para que não fosse encarado como um artefato sem implicações pedagógicas, vislumbrando-o como uma ferramenta de apoio para a prática avaliativa. Contudo, a utilização dos itens do formato impresso limitou a possibilidade de integrar itens com novos formatos e com uso de elementos suportados pela tecnologia, como animações, vídeos, sons, gráficos, objetos, entre outros. Além disso, os recursos tecnológicos devem dar abertura para formas informatizadas de pontuação, com a inclusão de correção de

automatizada de itens de resposta construída, contemplando a oferta de itens para avaliar a escrita, reivindicados na formação pelos(as) professores(as). Desse modo, os novos tipos de itens podem permitir aferição de habilidades mais complexas, comparadas às que estavam limitadas no instrumento impresso, constituindo aspectos para estudos posteriores.

Nestes termos, será fundamental que pesquisas vindouras possam focalizar o critério de seleção de itens, considerando os controles das taxas de exposição deles, visando melhorar a segurança do BI e diminuir o viés na estimação da proficiência em aplicações frequentes, além dos controles de conteúdo, aperfeiçoando a validade do teste.

Conjecturam-se igualmente investigações sobre a ampliação do BI do TAI da PB – Leitura, com mais itens da versão impressa, novos tipos e formatos de itens e itens ajustados às pessoas com deficiências. A ampliação do BI para itens novos exigirá a calibração desses itens, processo que deverá ser incorporado em módulo agregado ao algoritmo do TAI.

Ainda seria interessante abranger o teste de Matemática da PB, bem como avançar para a modificação do TAI com base em itens para que se tornasse um TAI de múltiplos estágios, que são testes que baseiam a adaptação somente depois de o(a) respondente ser submetido(a) a um agrupamento de itens. Considera-se que esses testes têm maior validade por possibilitar que o balanceamento de conteúdos seja realizado pelos especialistas.

Com a pesquisa desenvolvida, que inclusive possibilitou destacar algumas limitações do TAI construído e salientar perspectivas de continuidade do trabalho de investigação, considera-se que o TAI da PB – Leitura se mostrou efetivo em ser um ponto de apoio para a avaliação da proficiência em leitura de alunos(as) dos anos iniciais do ensino fundamental e promissor instrumento de avaliação educacional.

REFERÊNCIAS[56]

AL-AMRI, Saad. Computer-based testing vs. paper-based testing: a comprehensive approach to examing the comparability of testing modes. **Essex Graduate Student Papers in Language & Linguistic**, n. 10, p. 22-44, 2008.

ALAVARSE, Ocimar Munhoz. Desafios da avaliação educacional: ensino e aprendizagem como objetos de avaliação para a igualdade de resultados. **Cadernos Cenpec**, São Paulo, v. 3, n. 1, p. 135-153, jun. 2013.

ALAVARSE, Ocimar Munhoz; MELO, Wolney Candido de. Avaliação educacional e testes adaptativos informatizados (TAI): desafios presentes e futuros. In: BARBOSA, Alexandre F. (Coord.). **Pesquisa sobre o uso das tecnologias de informação e comunicação no Brasil:** TIC Educação 2012. São Paulo: Comitê Gestor da Internet no Brasil, 2013a. p. 103-112.

ALAVARSE, Ocimar Munhoz; MELO, Wolney Candido de. Educational evaluation and computerized adaptive testing (CAT): current and future challenges. In: BARBOSA, Alexandre F. (Coord.). **Survey on the use of information and communication technologies in Brazil:** ICT Education 2012. Translation by DB Comunicação. São Paulo: Comitê Gestor da Internet no Brasil, 2013b. p. 263-272.

ALAVARSE, Ocimar Munhoz. A organização do ensino fundamental em ciclos: algumas questões. **Revista Brasileira de Educação**, Rio de Janeiro, v. 14, n. 40, p. 35-50, jan./abr. 2009.

ALAVARSE, Ocimar Munhoz. Avaliações externas e seus efeitos. In: Mostra do CAEM 2015: 30 anos de formação continuada de professores de matemática, 2015, São Paulo. **Mostra do CAEM 2015: 30 anos de formação continuada de professores de matemática**. São Paulo: IME-USP, 2015. v. 1. p. 1-8.

ALAVARSE, Ocimar Munhoz; CATALANI, Érica Toledo. Alfabetização e TIC: os testes adaptativos informatizados (TAI) como recurso. In: COMITÊ GESTOR DA INTERNET NO BRASIL – CGI.br. **Pesquisa sobre o uso das tecnologias de informação e comunicação nas escolas brasileiras:** TIC Educação 2015. São Paulo: CGI.br, 2016a. p. 35-44.

ALAVARSE, Ocimar Munhoz; CATALANI, Érica Toledo. Literacy and ICT: computerized adpative testing as a resource. In: COMITÊ GESTOR DA INTERNET NO BRASIL – CGI.br. **Survey on the use of information and communication technologies in brazilians schools:** ICT in education. São Paulo: CGI.br, 2016b. p. 187-196.

ALAVARSE, Ocimar Munhoz; et al. Teste adaptativo informatizado como recurso tecnológico para alfabetização inicial. In: Séptima Conferencia Iberoamericana de Complejidad, Informática y Cibernética (CICIC 2017), 2017, Orlando, **Memorias.** Orlando: International Institute of Informatics and Systemics (IIIS), 2017, p. 165-169.

[56] De acordo com a Associação Brasileira de Normas Técnicas (ABNT NBR 6023).

ALAVARSE, Ocimar Munhoz et al. Teste adaptativo informatizado como recurso tecnológico para alfabetização inicial. **Revista Iberoamericana de Sistemas, Cibernética e Informática: RISCI**, v. 15, n. 3, p. 68-78, 2018a.

ALAVARSE, Ocimar Munhoz et al. O trabalho de formação para avaliação da alfabetização na perspectiva de um teste adaptativo informatizado. In: V Congresso Nacional de Avaliação em Educação (Conave): Da Educação Básica à Educação Superior: avaliação, formação de professores e direito à educação, 2018, Bauru, **Atas**. Bauru: UNESP/FC/Departamento de Educação, 2018b, (s.p).

ÁLVAREZ MENDEZ, Juan Manuel. **Avaliar para conhecer, examinar para excluir**. Tradução Magda Schwartzhaupt Chaves. Porto Alegre: Artmed, 2002.

ALMEIDA, Caroline Medeiros Martins de. **Prática educativa usando o sistema Siena para o ensino de ecologia no 6° ano do ensino fundamental.** 109 p. Dissertação (Mestrado em Ensino de Ciências e Matemática) – Universidade Luterana do Brasil, Canoas, 2014.

ABAD, Francisco José et al. Deterioro de parámetros de los ítems en tests adaptativos informatizados: estudio con eCAT. **Psicothema**, v. 22, n. 2, p. 340-347, 2010.

ABREU, Renata Cardoso Pires de. **Ensaio da Ferramenta DIA** – Diagnóstico e informação do aluno. 98 f. Dissertação (Mestrado em Ciências Computacionais) – Instituto de Matemática e Estatística, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2012.

AGUILAR, Gabriela; KAIJIRI, Kenji. Design overview of an adaptive computer-based assessment system. **Interactive Educational Multimedia**, n. 14, p. 116-130, April 2007.

ANDRADE, Dalton F.; VALLE, Raquel da C. Introdução à Teoria de Resposta ao Item: conceitos e aplicações. **Estudos em Avaliação Educacional**, v.18, p. 13-32, 1998.

ANDRADE, Dalton F.; TAVARES, Heliton R.; VALLE, Raquel da C**. Teoria da Resposta ao Item:** conceitos e aplicações. São Paulo: Associação Brasileira de Estatística, 2000.

ARAUJO, Joacy Victor Maia. **Teoria da resposta ao item em processo de decisão**. 68 p. Dissertação (Mestrado em Estatística) – Instituto de Ciências Exatas, Universidade de Brasília, Brasília, 2014.

BABCOCK, Ben; WEISS, David J. Termination criteria in Computerized Adaptive Tests: do variable-length CATs provide efficient and effective measurement? **Journal of Computerized Adaptive Testing**, v. 1, n. 1, Dec. 2012.

BAKER, Frank B. **The basics of Item Response Theory**. 2nd. ed. Washington: ERIC Clearinghouse on Assessment and Evaluation, 2001.

BARRADA, Juan Ramón. Tests adaptativos informatizados: una perspectiva general. **Anales de Psicología**, v. 28, n. 1, p. 289-302, ene. 2012.

BARRADA, Juan Ramóm et al. Estrategias de selección de ítems en un test adaptativo informatizado para la evaluación de inglés escrito. **Psicothema**, v. 18, n. 4, p. 828-834, 2006.

BARRADA, Juan Ramóm et al. Item selection rules in computerized adaptive testing: accuracy and security. **Methodology**, v. 5, n. 1, p. 7-17, 2009a.

BARRADA, Juan Ramón et al. **Test overlap rate and item exposure rate as indicators of test security in CATs**. Paper presented at the 2009 GMAC Conference on Compuretized Adaptive Testing. Disponível em: http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/cat09barrada.pdf. Acesso em: 17 nov. 2013.

BARRADA, Juan Ramón et al. A method for the comparison of item selection rules in Computerized Adaptive Testing. **Applied Psychological Measurement**, v. 34, n. 6, p. 438-452, 2010.

BARRADA, Juan Ramón et al. **Item bank disclosure in computerized adaptive testing: what makes an item selection rule safer?** Relatorio de pesquisa. Universidad Autónoma de Madrid, 2011. Disponível em: http://web.uam.es/becarios/jbarrada/papers/disclosure.pdf. Acesso em: 13 nov. 2013.

BARRADA, Juan Ramón; MAZUELA, Paloma; OLEA, Julio. Maximum information stratification method for controlling item exposure in Computerized Adaptive Testing. **Psicothema**, v. 18, n. 1, p. 156-159, 2006.

BARRADA, Juan Ramón; OLEA, Julio; PONSODA, Vicente. Methods for restricting maximum exposure rate in Computerized Adaptative Testing. **Methodology**, v. 3, n. 1, p. 14-23, 2007.

BARRADA, Juan Ramón; OLEA, Julio; PONSODA, Vicente; ABAD, Francisco José. Incorporating randomness to the Fisher information for improving item exposure control in CATS. **British Journal of Mathematical and Statistical Psychology**, v. 61, n. 2, p. 493-513, Nov. 2008.

BARRADA, Juan Ramón; OLEA, Julio; ABAD, Francisco José. Rotating item banks versus restriction of maximum exposure rates in Computerized Adaptive Testing. **The Spanish Journal of Psychology**, v. 11, n. 2, p. 618-625, 2008.

BARRADA, Juan Ramón; ABAD, Francisco José; VELDKAMP, Bernard P. Comparison of methods for controlling maximum exposure rates in Computerized Adaptive Testing. **Psicothema**, v. 21, n. 2, p. 313-320, 2009.

BARRADA, Juan Ramón; ABAD, Francisco José; OLEA, Julio. Optimal number of strata for the stratified methods in Computerized Adaptive Testing. **Spanish Journal of Psychology**, v. 17, e-48, p. 1-9, 2014.

BAYLARI, Ahmad; MONTAZER, G. A. Design a personlized e-learning system based on Item Response Theory and artificial neural network approach. **Expert Systems with Applications**, n. 36, p. 8013-8021, 2009.

BECHER, Ednei Luis. **Características do pensamento algébrico de estudantes do 1º ano do Ensino Médio.** 107 p. Dissertação (Mestrado em Ensino de Ciências e

Matemática), Universidade Luterana do Brasil, Canoas, 2009.

BEJAR, Isaac I. A validity-based approach to quality control and assurance of automated scoring. **Assessment in Education: Principles, Policy & Practice**, v. 18, n. 3, p. 319-341, Aug. 2011.

BJORNER, Jakob B. **Developing tailored instruments:** item banking and computerized adaptive assessment. Paper presentated at the conference Advances in Health Outcomes Measurement. Bethesda, Maryland, June 23-25, 2004.

BJORNER, Jakob B.; KOSINSKI, Mark; WARE JR, John E. Computerized adaptive testing and item banking. In: FAYERS, P.; HAYS, R. (Ed.). **Assessing quality of life in clinical trials**. 2nd ed. Los Angeles, CA: Oxford University Press, 2005. p. 95-112.

BOCK, R. Darrell; MISLEVY, Robert J. Adaptive EAP Estimation of ability in a microcomputer environment. **Applied Psychological Measurement**, v. 6, n. 4, p. 431-444, Sept. 1982.

BORSBOOM, Denny. **Conceptual issues in psychological measurement**. Amserdam: Universiteit van Amsterdam, 2003.

BORSBOOM, Denny. **Conceptual issues in contemporary psychometrics**. Amserdam: Cambridge University Press, 2005.

BLOOM, Benjamin.; HASTINGS, J. Thomas; MADAUS, George F. **Manual de avaliação formativa e somativa do aprendizado escolar**. Tradução de Lilian Rochlitz, Maria Cristina Fioratti Florez e Maria Eugênia Vanzolini. São Paulo: Pioneira, 1983. (Biblioteca Pioneira de Ciências Sociais). [Original 1971]

BRASIL. Lei nº 10,172, de 21 de março de 2005. Institui o Sistema de Avaliação da Educação Básica - SAEB, composto por dois processos de avaliação: a Avaliação Nacional da Educação Básica - ANEB, e a Avaliação Nacional do Rendimento Escolar - ANRESC. **Diário Oficial da União**, Brasília, DF, 22 mar. 2005. Seção 1.

BRASIL. Portaria nº 931, de 9 de janeiro de 2001. Aprova o Plano Nacional de Educação e dá outras providências. **Diário Oficial da União**, Brasília, DF, 10 jan. 2001. Seção 1.

BRASIL. Lei nº 11.274, de 6 de fevereiro de 2006. Altera a redação dos arts. 29, 30, 32 e 87 da Lei no 9.394, de 20 de dezembro de 1996, que estabelece as diretrizes e bases da educação nacional, dispondo sobre a duração de 9 (nove) anos para o ensino fundamental, com matrícula obrigatória a partir dos 6 (seis) anos de idade. **Diário Oficial da União**, Brasília, DF, 7 fev. 2006. Seção 1.

BRASIL. Portaria Normativa nº 10, de 24 de abril de 2007. Institui a Avaliação de Alfabetização “Provinha Brasil”. **Diário Oficial da União**, Brasília, DF, 26 abr. 2007. Seção 1.

BRASIL. Constituição (1988). Emenda Constitucional nº 59, de 11 de novembro de 2009. Acrescenta § 3º ao art. 76 do Ato das Disposições Constitucionais Transitórias para reduzir, anualmente, a partir do exercício de 2009, o percentual da Desvinculação das Receitas da União [...]. **Diário Oficial da União**, Brasília, DF, 12 nov. 2009. Seção 1.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Nacionais Anísio Teixeira (Inep). **Metodologia de validação e estruturação dos parâmetros psicométricos das edições de 2010 e 2011 da Provinha Brasil**. Responsável Técnico Cácio Fabricio Gomes da Rocha. Brasília, 2011a.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Nacionais Anísio Teixeira (Inep). **Documento Técnico B:** material psicométrico para oficina de interpretação de escalas da Provinha Brasil (Leitura). Responsável Técnico Adriano Ferreti Borgatto. Brasília, [2011b].

BRASIL. Portaria nº 867, de 4 de julho de 2012. Institui o Pacto Nacional pela Alfabetização na Idade Certa e as ações do Pacto e define suas diretrizes gerais. **Diário Oficial da União**, Brasília, DF, 5 jul. 2012. Seção 1.

BRASIL. Ministério da Educação. Secretaria da Educação Básica (SEB). Instituto Nacional de Estudos e Pesquisas Anísio Teixeira (Inep). **Guia de elaboração de itens Provinha Brasil 2012**. Brasília, 2012b. Disponível em: < http://download.inep.gov.br/educacao_basica/provinha_brasil/documentos/2012/guia_elaboracao_itens_provinha_brasil.pdf>. Acesso em: 30 jul. 2016.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Nacionais Anísio Teixeira (Inep). **Relatório das estatísticas e pedagógicas dos itens (subgrupo 1):** 1º Pré-teste de itens do Banco Nacional de Itens (BNI) 2012. Cespe-Unb e Cengranrio, Responsável Técnico: Rolf Stöller Arruda. Brasília, 2012c.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Nacionais Anísio Teixeira (Inep). **PD8 (subgrupo 1) Análise clássica e TRI dos itens:** Pré-teste 2013.2 do Banco Nacional de Itens do Inep. Cespe-Unb e Cengranrio, Responsáveis Técnicos Haydée Werneck Poubel e Rolf Stöller Arruda. Brasília, 2013a.

BRASIL. Portaria nº 482, de 7 de junho de 2013. Dispõe sobre o Sistema de Avaliação da Educação Básica. **Diário Oficial da União**, Brasília, DF, 10 jun. 2013b. Seção 1.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Nacionais Anísio Teixeira (Inep). **PD8 (subgrupo 1) Análise clássica e TRI dos itens:** Pré-teste 2014.1 do Banco Nacional de Itens do Inep. Cespe-Unb e Cengranrio, Responsáveis Técnicos: Haydée Werneck Poubel e Rolf Stöller Arruda. Brasília, 2014.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Nacionais Anísio Teixeira (Inep). Provinha Brasil: avaliando a alfabetização. **Guia de correção e interpretação de resultados:** Leitura e Matemática, teste 1, 2015. Brasília: Ministério da Educação, Instituto Nacional de Estudos e Pesquisas Nacionais Anísio Teixeira (Inep), 2015a.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Nacionais Anísio Teixeira (Inep). **Guia de correção e interpretação de resultados:** Provinha Brasil – Leitura, teste 2, 2015. Brasília: Ministério da Educação, Instituto Nacional de Estudos e Pesquisas Nacionais Anísio Teixeira (Inep), 2015b.

BRASIL. Portaria nº 387, de 1 de setembro de 2015. Estabelece o inciso III do art. do art. 1 do Decreto nº 6.317, de 20 de dezembro de 2007, e tendo em vista o disposto na LEI Nº 13.005, DE 25 DE JUNHO DE 2014, que aprova o plano nacional de educação - PNE. **Diário Oficial da União**, Brasília, DF, 2 set. 2015c. Seção 1.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Nacionais Anísio Teixeira (Inep). **Guia de interpretação de resultados:** Provinha Brasil – Leitura, teste 2, 2016. Brasília: Ministério da Educação, Instituto Nacional de Estudos e Pesquisas Nacionais Anísio Teixeira (Inep), 2016a.

BRASIL. Ministério da Educação. Instituto Nacional de Estudos e Pesquisas Nacionais Anísio Teixeira (Inep). **Caderno do aluno:** Provinha Brasil – Leitura, teste 2, 2016. Brasília: Ministério da Educação, Instituto Nacional de Estudos e Pesquisas Nacionais Anísio Teixeira (Inep), 2016b.

BRASIL.Resolução CNE/CP nº 2, de 22 de dezembro de 2017. Institui e orienta a implantação da Base Nacional Comum Curricular, a ser respeitada obrigatoriamente ao longo das etapas e respectivas modalidades no âmbito da Educação Básica. **Diário Oficial da União**, Brasília, DF, 22 dez. 2017. Seção 1.

BRASIL. Decreto nº 9.432, de 29 de junho de 2018. Regulamenta a Política Nacional de Avaliação e Exames da Educação Básica. **Diário Oficial da União**, Brasília, DF, 2 jul. 2018. Seção 1.

CARDINET, Jean. **Avaliar é medir?** Porto: Edições Asa, 1993. (Práticas Pedagógicas).

CASTRO, Natália Fontes Caputo de. **Tradução e adaptação transcultural do domínio fadiga do Patient-Reported-Outcomes Measurement Information System – PROMIS® – para a língua portuguesa**. 81 f. Dissertação (Mestrado em Ciências da Saúde) – Faculdade de Medicina, Universidade Federal de Uberlândia, Uberlândia, 2013.

CATALANI, Érica Maria Toledo; TATAGIBA, Alessandro Borges. Provinha Brasil: desafios e perspectivas para a apropriação pedagógica dos resultados. In: **ASSOCIAÇÃO BRASILEIRA DE AVALIAÇÃO EDUCACIONAL.** VIII Reunião da Associação Brasileira de Avaliação Educacional: avaliação de larga escala no Brasil: ensinamentos, aprendizagens e tendências: anais. Florianópolis: Abave, 2015. p. 493-496.

CHAJEWSKI; Michael; LEWIS, Charles. **Optimizing item exposure control algorithms for polytomous computerized adaptive tests with restricted item banks.** Paper presented at the 2009 GMAC Conference on Compuretized Adaptive Testing. Disponível em: http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/cat09chajewski.pdf. Acesso em: 17 nov. 2013.

CHANG, Hua-Hua. Making computerized adaptive testing diagnostic tools for schools. In: LISSITZ, Robert W.; JIAO, Hong (Ed.). **Computers and their impact on state assessments**: recent history and predictions for the future. Charlotte, NC: Information Age, 2012. p. 195-226.

CHANG, Hua-Hua; VAN DER LINDEN, Wim J. Optimal stratification of Item Pools in a-Stratified Computerized Adaptive Testing. **Applied Psychological Measurement**, v. 27, n. 4, p. 262-274, July 2003.

CHANG, Hua-Hua; YING, Zhiliang. A global information approach to computerized adaptive testing. **Applied Psychological Measurement**, v. 20, n. 3, p. 213-229, Sep. 1996.

CHANG, Hua-Hua. Psychometrics behind computerized adaptive testing. **Psychometrika,** p. 1-20, 6 Feb. 2014.

CHANG, Hua-Hua; ZHANG, Jinming. Hypergeometric family and item overlap rates in computerized adaptive testing. **Psychometrika**, v. 67, n. 3, p. 387-398, Sep. 2002.

CHANG, Shun-Wen; ANSLEY, Timothy N. A comparative study of item exposure control methods in Computerized Adaptive Testing. **Journal of Educational Measurement**, v. 40, n. 1, p. 71-103, Spring 2003.

CHENG, Ying. When Cognitive Diagnosis meets Computerized Adaptive Testing: CD-CAT. **Psychometrika**, v. 74, n. 4, p. 619-632, Dec. 2009.

CHENG, Ying. Improving Cognitive Diagnostic Computerized Adaptive Testing by balancing attribute coverage: the Modified Maximum Global Discrimination Index Method. **Educational and Psychological Measurement**, v. 70, n. 6, p. 902-913, 2010.

CHEN, Shu-Ying; ANKENMANN, Robert D.; SPRAY, Judith. The relationship between item exposure and test overlap in computerized adaptive testing. **Journal of Educational Measurement**, v. 40, n. 2, p. 129-145, Summer 2003.

CHEN, Deng-Jyi; LAI, Ah-Fur; MAO, Chia-Chi. **The analysis of response patterns on IRT ability estimation methods in computarized adaptive test.** Paper presented at the Seventh IEEE International Conference on Advanced Learning Technologies, 2007.

CHEN, Shu-Ying; DOONG, Shing-Hwang. **Predicting item exposure parameters in computerized adaptive testing**. Paper presented at the Annual Meeting of the American Educational Research Association. Chicago. 2003.

CHEN, Shu-Ying; LEI, Pui-Wa. Controlling item exposure and test overlap in Computerized Adaptive Testing. **Applied Psychological Measurement**, v. 29, n. 3, p. 204-217, May 2005.

CHEN, Yunxiao; LIU, Jingchen; YING, Zhiliang. Online item calibration for Q-Matriz in CD-CAT**. Applied Psychological Measurement**, v. 39, n. 1, p. 5-15, 2015.

CIZEK, Gregory J.; BUNCH, Michael B. **Standard setting:** a guide to establishing and evaluating performance standards on tests. Thousand Oaks: Sage, 2007.

CLARES LÓPEZ, José. Propuesta de desarrollo de test informatizado adaptándolo a las respuestas del usuario. **Pixel-Bit. Revista de Medios y Educación**, n. 31, p. 19-30, ene. 2008.

CLARIANA, Roy; WALLACE, Patricia. Paper-based versus computer-based assessment: key factors associated with the test mode effect. **British Journal of Educational Technology**, v. 33, n. 5, p. 593-602, 2002.

COSTA, Denise Reis. **Métodos estatísticos em testes adaptativos informatizados**. 120 p. Dissertação (Mestrado) – Instituto de Matemática, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2009.

COSTA, Patrícia; FERRÃO, Maria Eugénia. On the complementarity of classical test theory and item response models: item difficulty estimates and computerized adaptive testing. **Ensaio: Avaliação e Políticas Públicas em Educação**, Rio de Janeiro, v. 23, n. 88, p. 593-610, jul./set. 2015.

COUTO, Gleiber; PRIMI, Ricardo. Teoria de resposta ao item (TRI): conceitos elementares dos modelos para itens dicotômicos. **Boletim de Psicologia**, São Paulo, v. 61, n. 134, p. 1-15, jun. 2011.

CRAHAY, Marcel. **Poderá a escola ser justa e eficaz?**: da igualdade das oportunidades à igualdade dos conhecimentos. Tradução de Vasco Farinha. Lisboa: Instituto Piaget, 2002. (Horizontes Pedagógicos, 92). [Original 2000]

DAVEY, Tim. **A guide to computer adaptive testing systems**. Washington, DC: Technical Issues in Large-Scale Assessment (TILSA): State Collaborative on Assessment and Student Standards (SCASS). Nov. 2011.

DAVEY, Tim; PITONIAK, Mary J. Designing computerized adaptive tests. In: DOWNING, Steven M; HALADYNA, Thomas M. (Ed.). **Handbook of test development**. Mahwah, NJ: Lawrence Erlbaum, 2006. p. 543-473.

DE AYALA, R. J. **The Theory and Practice of Item Response Theory**. New York: The Guilford Press, 2009. (Methodology in the Social Sciences).

DESTRO, Bruno de Jesus; MENEGHETTI, Douglas De Rizzo. **Desenvolvimento de um sistema de aplicação de testes Informatizados com conteúdo multimídia**. 75 f. Monografia (TCC) – Faculdade de Tecnologia e Termodinâmica, São Bernardo do Campo, 2012.

DODD, Barbara G. The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. **Applied Psychological Measurement**, v. 14, n. 4, p. 355-366, Dec. 1990.

DOLAN, Robert P.; BURLING, Kelly S. Computer-based testing in higher education. In: SECOLSKY, Charles; DENISON, Brian (Ed.). **Handbook on measurement, assessment and evaluation in higher education**. 2nd ed. New York: Routledge, 2018. p. 370-384.

DOLAN, Robert P. et al. Applying principles of universal design to test delivery: the effect of computer-based read-aloud on test performance oh high school students with learning disabilities. **The Journal of Technology, Learning and Assessment**, v. 3, n. 7, p. 1-32, Feb. 2005.

DORANS, Neil J. Scaling and equating. In: WAINER, Howard et al. **Computerized adaptive testing**: a primer. 2nd ed. New York: Routledge, 2014. p. 135-158. [Original 2000]

EMBRETSON, Susan E.; REISE, Steven Paul. **Item Response Theory for Psychologists Multivariate**. New Jersey: Lawrence Erlbaum, 2000.

EHLERS, Ricardo Sanders. Introdução a Inferência Bayesiana. 7 ed. Departamento de Estatística, Universidade Federal do Paraná, 2007. Disponível em:<http://conteudo.icmc.usp.br/pessoas/ehlers/bayes/bayes.pdf> Acesso em 17 abr 2019.

ESTEBAN, M. T. Provinha Brasil: desempenho escolar e discursos normativos sobre a infância. **Sísifo – Revista de Ciências da Educação**, n. 9, p. 47-55, maio/ago. 2009.

EGGEN, Theo J .H. M. **Overexposure and underexposure of items in Computerized Adaptive Testing**. Arnhem: Citogroep, januari 2001. (Measurement and Research Department Reports, 2001).

EGGEN, Theo J. H. M.; STRAETMANS, G. J. J. M. Computerized adaptive testing for classifying examinees into three categories. **Educational and Psychological Measurement**, v. 60, n. 5, p. 713-734, Oct. 2000.

EL-ALFY, El-Sayed M.; ABDEL-AAL, Radwan E. **Construction and analysis of educational tests using abductive machine learning**. Dhahran: College of Computer Sciences and Engineering: King Fahd University of Petroleum and Minerals, [c. 2008].

FERNANDES, Cláudia de Oliveira; FREITAS, Luiz Carlos de. **Indagações sobre currículo**: currículo e avaliação. Brasília: Ministério da Educação, Secretaria de Educação Básica, 2007.

FERNANDES, Domingos. **Avaliar para aprender**: fundamentos, práticas e políticas. São Paulo: Editora Unesp, 2009.

FERRÃO, Maria Eugénia; PRATA, Paula. **Item Response Models in Computerized Adaptive Testing**: a simulation study. In: MURGANTE, B. et al. (Ed.). ICCSA 2014: part III. [S.l.]: 2014. p. 552-565.

FLETCHER, P. R. **Introdução à Teoria de Características Latentes e Modelos de Resposta ao Item**. Rio de Janeiro: Escola Nacional de Ciências Estatísticas, fev. 2000. Disponível em: http://www.avaliaeducacional.com.br/referencias/arquivos/Philip%20-%20Introducao%20a%C2%A0%20Teoria%20de%20Caracteristicas%20Latentes%20e%20Modelos%20de%20Resposta%20ao%20Item.pdf. Acesso em: 16 mar. 2014.

FOSTER, David. Testing technology and its effects on test security. In: DRASGOW, Fritz (Ed.). **Technology and testing:** improving educational and psychological measurement. New York: Routledge, 2015. (NCME Applications of Educational Measurement and Assessment). p. 235-254.

FRIES, James F. et al. Item Response Theory, Computerized Adaptive Testing, and PROMIS: assessment of physical function. **The Journal of Rheumatology**, v. 41, n. 1, p. 153-158, 2014.

GALVÃO, Ailton Fonseca. **Um Modelo Inteligente para Seleção de Itens em Testes Adaptativos Computadorizados**. 79 p. Dissertação (Mestrado acadêmico em Ciências da computação) – Universidade Federal de Juiz de Fora, Juiz de Fora, 2013.

GARCÍA JIMÉNEZ, Eduardo; GIL FLORES, Javier; RODRÍGUEZ GÓMEZ, Gregorio. La evaluación de tests adaptativos informatizados. **Revista Electrónica de Investigación y Evaluación Educativa**, v. 4, n. 2, 1998.

GATTI, Bernardete Angelina et al. Formação de professores para o ensino fundamental: instituições formadoras e seus currículos. **Estudos & Pesquisas Educacionais**, São Paulo, n. 1, p. 95-138, 2010.

GATTI, Bernardete Angelina. Avaliação: contexto, história e perspectivas. **Olh@res**, Guarulhos, v. 2, n. 1, p. 08-26, mai. 2014.

GEISINGER, Kurt F. Commentary on chapters 8-11: technology and test administration: the search for validity. In: DRASGOW, Fritz (Ed.). **Technology and testing:** improving educational and psychological measurement. New York: Routledge, 2015. (NCME Applications of Educational Measurement and Assessment). p. 255-259.

GEORGIADOU, Elissavet; TRIANTAFILLOU, Evangelos; ECONOMIDES, Anastasios. A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. **The Journal of Technology, Learning, and Assessment**, v. 5, n. 8, May 2007.

GERHARDT, Tatiana Engel; SILVEIRA, Denise Tolfo (Coord). **Métodos de pesquisa.** Universidade Aberta do Brasil – UAB/UFRGS, Curso de Graduação Tecnológica – Planejamento e Gestão para o Desenvolvimento Rural da SEAD/UFRGS. Porto Alegre: UFRGS, 2009.

GOLINO, Hudson F.; GOMES, Cristiano Mauro Assis. Teoria da medida e o Modelo Rasch. In: GOLINO, Hudson F. et al. (Org.). **Psicometria contemporânea**: compreendendo os Modelos Rasch. São Paulo: Casa do Psicólogo, 2015. p. 13-45.

GONZÁLEZ BETANZOS, Fabiola. **Detección del Funcionamiento Diferencial del Ítem em Test Adaptativos Inormatizados**. 75 f. Tesis (Doctoral) – Faculdad de Psicologia. Departaamento de Psicologia Social y Metodologia.Universidad Autónoma de Madrid. Madrid, 2011.

GRAUDINA, Vita; GRUNDSPENKIS, Janis. **Conceptual model for ontology-based adaptive assessment system**. 3rd E-Learning Conference, Coimbra, 7-8 Sep. 2006.

HAGUETTE, T. M. F. **Metodologias qualitativas na sociologia**. 6 ed. Petrópolis: Vozes, 1999. 224p.

HAMBLETON, Ronald K.; JONES, Russell W.; ROGERS, H. Jane. Influence of Item Parameter Estimation Errors in Test Development. **Journal of Educational Measurement**, v. 30, n. 2, p. 143-155, 1993.

HAMBLETON, R. K.; SWAMINATHAN, H. **Item Response Theory:** principles and applications. Boston: Kluwer.Nijhoff, 1985. (Evaluation in Education and Human Services).

HAN, Kyung (Chris) Tyek. Components of the item selection algorithm in computerized adaptive testing. **Journal of Educational Evaluation for Health Professions**, v. 15, n. 7, p. 1-13, 2018.

HARMS, Michael; ADAMS, Jeremy. **Usability and design considerations for computer-based learning and assessment.** Paper presented at the March 2008 Meeting of the American Educational Research Association (AERA).

HAU, Kit-Tai; CHANG, Hua-Hua. Item selection in computerized adaptive testing: should more discriminating items be used first? **Journal of Educational Measurement**, v. 38, n. 3, p. 249-266, Fall 2001.

HOFFMANN, Jussara Maria Lerch. **Avaliação mediadora:** uma prática em construção da pré-escola à universidade. 20. ed. Porto Alegre: Mediação, 2003.

HOMA, Agostinho Iaqchan Ryokiti. ***E-learning* com análise combinatória**. 106 f. Dissertação (Mestrado em Ensino de Ciências e Matemática) – Universidade Luterana do Brasil, Canoas, 2012.

HO, Rong-Guey; YEN, Yung-Chin. Design and evaluation of an XML-based platform-independent computerized Adaptive Testing System. **IEEE Transactions on Education**, v. 48, n. 2, p. 230-237, May 2005.

HUEBNER, Alan. An overview of recent developments in cognitive diagnostic Computer Adaptive Assessments. **Practical Assessment, Research & Evaluation**, v. 15, n. 3, p. 1-7, Jan. 2010.

KARKEE, Thakur; KIM, Dong-In; FATICA, Kevin. **Comparability study of online and paper and pencil tests using modified internally and externally matched criteria**. Paper presented at the Annual Meeting of the American Educational Research Association (AERA). Denver, CO. April 29-May 4, 2010.

KENG, Leslie. **A comparison of the performance of Testlet-Based Computer Adaptive Tests and Multistage Tests.** 229 p. Doctoral dissertation (Doctorade of Philosophy) – Faculty of the Graduate School, University of Texas at Austin, Austin, 2008.

KENG, Leslie et al. **A comparison of item and testlet selection procedures in Computerized Adaptive Testing**. [S.l.n.]: c. 2010.

KIM, Do-Hong; HUYNH, Huynh. Comparability of computer and paper-and-pencil versions of Algebra and Biology assessments. **The Journal of Technology, Learning and Assessment**, v. 6, n. 4, p. 1-30, Dec. 2007.

KIM, Sooyeon; MOSES, Tim; YOO, Hanwook. A comparison of IRT proficiency estimation methods under adaptive multistage testing. **Journal of Educational Measurement**, v. 52, n. 1, p. 70-79, Spring 2015a.

KIM, Sooyen; MOSES, Tim; YOO, Hanwook Henry. **Effectiveness of Item Response Theory (IRT) proficiency estimation methods under adaptive multistage testing.** Princeton, NJ: Educational Testing Service, June 2015b. (Research Report ETS RR-15-11).

KIM KANG, Gyenamkim; WEISS, David J. **Comparison of computerized adaptive testing and classical methods for measuring individual change**. Paper presented at the Item Calibration and Special Applications Paper Session, 2007. GMAC Conference on Computerized Adaptive Testing, June 7, 2007.

KIMURA, Tetsuo. The impacts of computer adaptive testing from a variety of perspectives. **The Journal of Education Evaluation for Health Professions**, v. 14, n. 8, p. 1-5, May 2017.

KINGSBURY, G. Gage; HOUSER, Ronald L. **A comparison of achievement level estimates from computerized adaptive testing and paper-and-pencil testing**. A paper presented to the Annual Meeting of the American Educational Research Association. New Orleans, LA. April 9, 1988.

KINGSBURY, G. Gage; HOUSER, Ronald L. Assessing the utility of Item Response Models: Computerized Adaptive Testing. **Educational Measurement:** Issues and Practice, v. 12, n. 1, p. 21-27, Spring 1993.

KINGSBURY, G. Gage; ZARA, Anthony R. Procedures for selecting items for Computerized Adaptive Tests. **Applied Measurement in Education**, v. 2, n. 4, p. 359-375, 1989.

KINGSBURY, G. Gage; WEISS, David J. A Comparison of IRT-Based Adaptive

Mastery Testing and a Sequential Mastery Testing Procedure. In: WEISS, David J. **New Horizons in Testing**: Latent Trait Test Theory and Computerized Adaptive Testing. 2nd ed. New York: Academic Press, Inc, 1983. p. 257-283.

KLEIN, Ruben. Utilização da Teoria da Resposta ao Item no Sistema Nacional de Avaliação da Educação Básica (Saeb). **Ensaio: Avaliação e Políticas Públicas em Educação**, Rio de Janeiro, v. 11, n. 40, p. 283-296, jul./set. 2003.

KLEIN, Rubem. Alguns aspectos da teoria de resposta ao item relativos à estimação das proficiências. **Ensaio: Avaliação e Políticas Públicas em Educação**, Rio de Janeiro, v. 21, n. 78, p. 35-56, jan./mar. 2013.

KRANTZ, David H. et al. **Foundations of measurement**: additive and polynomial representations. v. I. New York: Academic Press, 1971.

KREITZBERG, Charles B.; STOCKING, Martha L.; SWANSON, Len. Computerized adaptive testing: principles and directions. **Computers and Education**, v. 2, n. 4, p. 319-329, 1978.

LI, Xiaoping et al. The design of adaptive test paper composition algorithm based on the Item Response Theory. **Information Technology and Artificial Intelligence Conference (ITAIC), 2011 6th IEEE Joint Internacional**, v. 2, p. 157-159, Aug. 2011.

LÓPEZ-CUADRADO, Javier; PÉREZ, Tomás A.; ARMENDARIZ, Ana Jesús. Evaluación mediante tests: ¿Por qué no usar el ordenador? **Revista Iberoamericana de Educación**, v. 36, n. 11. p. 1-15, oct. 2005.

LORD, Frederic M. A broad-range tailored test of verbal ability. **Applied Psychological Measurement**, v. 1, n. 1, p. 95-100, Winter 1977.

LORD, Frederic M. **Applications of item response theory to practical testing problems**. Hillsdale: Lawrence Erlbaum, 1980.

LORD, Frederic M.; NOVICK, Melvin R. **Statistical theories of mental test scores**. Reading, MA: Addison-Wesley, 1968. (The Addison-Wesley Series in Behavioral Science. Quantitative Methods).

LOZZIA, Gabriela; ATTORRESI, Horacio. Especificación del algoritmo para un Test Adaptativo Informatizado de analogías verbales. **Summa Psicológica UST**, v. 9, n. 2, p. 15-23, 2009.

LU, Peng; CONG, Xiao. The research on Computerized Adaptive Testing. **Journal of Physics:** Conference Series, 710, p. 1-10, 2016.

LUECHT, Richard M. Computer-based and computer-adaptive testing. In: SIMON, Marielle; ERCIKAN, Kadriye; ROUSSEAU, Michel (Ed.). **Improving large-scale assessment in education: theory, issues and practice**. New York: Routledge, 2013. p. 62-84.

LUECHT, Richard M. Computer-based test delivery models, data and operational implementation issues. In: DRASGOW, Fritz (Ed.). **Technology and testing:** improving educational and psychological measurement. New York: Routledge, 2015. (NCME Applications of Educational Measurement and Assessment). p. 179-205.

LUECHT, Richard M.; SIRECI, Stephen G. **A review of models for Computer-Based Testing.** New York: College Board, 2011. (Research Report 2011-12).

LUCKESI, Cipriano Carlos. **Avaliação da aprendizagem escolar**: estudos e proposições. 19 ed. São Paulo: Cortez, 2008.

LUCKESI, Cipriano Carlos. **Avaliação em educação**: questões epistemológicas e práticas. São Paulo: Cortez, 2018.

LUKAS MUJIKA, José Francisco; SANTIAGO ETXEBARRÍA, Karlos. **Evaluación educativa**. 2. ed. Madrid: Alianza, 2009.

MAGIS, David et al. A general framework and an R package for the detection of dichotomous differential iten functioning. **Behavior Research Methods**, v. 42, n. 3, p. 847-862, Aug. 2010.

MAGIS, David; MAHALINGAM, Vaishali. Computerized adaptive testing. In: SILVA, Marjorie Cristina Rocha da et al. (Org.). **Aplicações de métodos estatísticos avançados à avaliação psicológica e educacional:** com ilustrações em diferentes softwares estatísticos. São Paulo: Vetor, 2015. p. 239-256.

MAGIS, David; RAÎCHE, Gilles. Random Generation of Response Patterns under Computerized Adaptive Testing with the R Package catR. **Journal of Statistical Software**, v. 48, n. 8, p. 1-31, May 2012.

MAGIS, David; YAN, Duanli; VON DAVIER, Alina A. **Computerized Adaptive and Multistage Testing with R**: using Packages catR and mstR. Cham: Springer, 2017.

MAHALINGAM, Vaishali; MAGIS, David. Computer adaptive testing using Concerto. In: SILVA, Marjorie Cristina Rocha da et al. (Org.). **Aplicações de métodos estatísticos avançados à avaliação psicológica e educacional:** com ilustrações em diferentes softwares estatísticos. São Paulo: Vetor, 2015. p. 211-238.

MAIA JÚNIOR, Antonio Geraldo Pinto. **Uso do tempo de resposta para melhorar a convergência do algoritmo de Testes Adaptativos Informatizados.** 69 p. Dissertação (Mestrado em Estatística) – Instituto de Ciências Exatas, Universidade de Brasília, Brasília, 2015.

MANSEIRA, Paulo Rogério Pires; MISAGHI, Mehran. **Proposta de ferramenta para uso abrangente de Testes Computadorizados na educação a distância**. Trabalho apresentado no III Congresso Brasileiro de Engenharia de Produção. Ponta Grossa, PR. 04 a 06 dezembro 2013.

MARTIN, Andrew J.; LAZENDIC, Goran. Computer-adaptive testing: implications for students' achievement, motivation, engagement, and subjective test experience. **Journal of Educational Psychology**, v. 110, n. 1, p. 27-45, 2018.

MARTIN, Romain. Le testing adaptatif par ordinateur dans la mesure en éducation: potentialités et limites. **Psychologie et Psychométrie**, v. 24, n. 2-3, p. 89-116, 2003.

MARTÍN-FERNÁNDEZ, Manuel et al. A multistage adaptive test of fluid intelligence. **Psicothema**, v. 28, n. 3, p. 346-352, 2016.

MÁXIMO, Luis Fernando. **A efetividade de feedbacks informatizados sobre a autoregulação da aprendizagem em cursos a distância**: um estudo de caso na área da computação.150 p. Tese (Doutorado em Informática na Educação) – Centro de Estudos Interdisciplinares em Novas Tecnologias da Educação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

MELO, Wolney Candido de. **Erros de medida da Prova Brasil 2013 e sua influência no Ideb das escolas da Rede Municipal de ensino de São Paulo**: um estudo sobre os anos finais do Ensino Fundamental e seus desdobramentos para as políticas educacionais. 217 p. Tese (Doutorado em Educação) – Faculdade de Educação da Universidade de São Paulo, São Paulo, 2017.

MENEGHETTI, Douglas De Rizzo. **Metodologia de seleção de itens em testes adaptativos informatizados baseada em agrupamento por similaridade.** 96 f. Dissertação (Mestrado) – Centro Universitário da FEI, São Bernardo do Campo, 2015.

MICOTTI, Maria Cecília de Oliveira. A avaliação do ensino e do aprendizado de língua portuguesa nas séries iniciais da escola fundamental. In: **AVALIAÇÕES da educação básica em debate:** ensino e matrizes curriculares de referência das avaliações em larga escala. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), 2013. p. 175-194,

MILLS, Craig N. et al. (Ed.). **Computer based testing:** building the foundation for future assessments. Mahwah, NJ: Lawrence Erlbaum, 2002.

MISLEVY, Robert J. et al. Psychometrics and game-based assessment. In: DRASGOW, Fritz (Ed.). **Technology and testing:** improving educational and psychological measurement. New York: Routledge, 2015. (NCME Applications of Educational Measurement and Assessment). p. 23-48.

MOLINA, M. Teresa López-mezquita. **La evaluación de la competência léxica:** tests de vocabulario.su fiabilidad y validez. 1076 p. Tese (Doutorado em Filología Inglesa) – Facultad de Filosofía y Letras, Universidad de Granada, Granada, 2005.

MOITA, Pedro Miguel da Silva. **Avaliação adaptativa em dispositivos móveis das habilidades cognitivas preditoras do desenvolvimento de leitura em crianças.** 132 f. Dissertação (Mestrado em Tecnologias e Metodologias em *E-learning*) – Instituto de Educação da Faculdade de Ciências, Universidade de Lisboa, Lisboa, 2013.

MORAIS, Artur Gomes de. **Sistema de Escrita Alfabética**. São Paulo: Melhoramentos, 2012. (Como Eu Ensino).

MORAIS, Artur Gomes de; LEAL, Telma Ferraz; PESSOA, Ana Cláudia Rodrigues Gonçalves. O ensino da língua portuguesa no ciclo de alfabetização e sua avaliação pela Provinha Brasil. In: **AVALIAÇÕES da educação básica em debate:** ensino e matrizes curriculares de referência das avaliações em larga escala. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), 2013.p. 153-174.

MOREIRA JUNIOR, Fernando de Jesus. **Sistemática para implantação de testes adaptativos informatizados baseados na Teoria da Resposta ao Item**. 334 p. Tese (Doutorado em Engenharia de Produção) – Centro Tecnológico, Universidade Federal de Santa Catarina, Florianópolis, 2011.

MOREIRA JUNIOR, Fernando de Jesus et al. Algoritmo de um teste adaptativo informatizado com base na teoria da resposta ao item para a estimação da usabilidade de sites de e-commerce. **Produção**, v. 23, n. 3, p. 525-536, jul./set. 2013.

MORENO, Kathleen E.; SEGALL, Daniel O.; HETTER, Rebecca D. The use of Computarized Adaptive Testing in the Military. In: DILLON, Ronna F. (Ed.). **Handbook on testing.** Westport, CT: Greenwood Press, 1997. p. 204-219.

MUÑIZ, José; HAMBLETON, Ronald K. Evaliación psicométrica de los tests informatizados. In: OLEA, Julio; PONSODA, Vicente; PRIETO, Geraldo. (Ed.). **Tests informatizados:** fundamentos y aplicaciones. Madrid: Pirámide, 1999. p. 23-43.

MUÑIZ, José. **Introducción a la Teoría de Respuesta a los Ítems**. Madrid: Pirámide, 1997. (Psicología).

NEVO, David. Avaliação por diálogos: uma contribuição possível para o aprimoramento escolar. In: TIANA, Alejandro (Coord.). **Anais do Seminário Internacional de Avaliação Educacional**, 1 a 3 de dezembro de 1997. Tradução de John Stephen Morris. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais (Inep), 1998. p. 89-97.

NEVO, David. **Evaliación basada en el centro:** un diálogo para la mejora educativa. Traducción Maria Serrano Bericat. Bilbao: Mensajero, 1997.

NUNES, Carlos Henrique Sancineto da Silva et al. Testes adaptativos computadorizados – CAT. In: SILVA, Marjorie Cristina Rocha da et al. (Org.). **Aplicações de métodos estatísticos avançados à avaliação psicológica e educacional:** com ilustrações em diferentes softwares estatísticos. São Paulo: Vetor, 2015. p. 37-76.

OLEA, Julio; PONSODA, Vicente; PRIETO, Gerardo (Ed.). **Tests informatizados:** fundamentos y aplicaciones. Madrid: Pirámide, 1999. (Psicología).

OLEA, Julio et al. Investigación en tests adaptativos informatizados. In: OLEA, Julio; PONSODA, Vicente; PRIETO, Gerardo (Ed.). **Tests informatizados:** fundamentos y aplicaciones. Madrid: Pirámide, 1999. p. 163-188.

OLEA, Julio et al. Un test adaptativo informatizado para evaluar el conocimiento de inglés escrito: diseño y comprobaciones psicométricas. **Psicothema**, v. 16, n. 3, p. 519-525, 2004.

OLEA, Julio et al. eCAT-Listening: design and psychometric properties of a computerized adaptive test on English Listening. **Psicothema**, v. 23, n. 4, p. 802-807, 2011.

OLEA, J.; PONSODA, V. **Tests Adaptativos Informatizados**. Madrid: Ediciones UNED, 2003.

OLEA, Julio; ABAD, Francisco J.; BARRADA, Juan R. Tests informatizados y otros nuevos tipos de tests. **Papeles del Psicólogo**, v. 31, n. 1, p. 94-107, 2010.

OLEA, Julio; PONSODA, Vicente. **Tests adaptativos informatizados.** Madrid: edición de los autores, 2002. Disponível em: https://www.researchgate.net/profile/Julio_Diaz5/publication/265040034_TEST_ADAPTATIVOS_INFORMATIZADOS/links/54e481630cf2dbf60696bbb5/TEST-ADAPTATIVOS-INFORMATIZADOS.pdf. Acesso em: 24 set. 2018.

OLIVEIRA, Cassandra Melo. **Construção e busca de evidências de validade de um banco de itens de personalidade para testagem adaptativa desenvolvido a partir dos princípios do Desenho Universal**.173 f. Tese (Doutorado em Psicologia) – Centro de Filosofia e Ciências Humanas, Universidade Federal de Santa Catarina, Florianópolis, 2017.

OLIVEIRA, Leandro Henrique Mendonça de. **Testes Adaptativos Sensíveis ao Conteúdo do Banco de Itens:** Uma Aplicação em Exames de proficiência em Inglês para Programas de Pós-Graduação. 220 p. Dissertação (Mestrado em Ciências da Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2002.

OZTURK, Nagihan Boztunc; DOGAN, Nuri. Investigating item exposure control methods in Computerized Adaptive Testing. **Educational Sciences: Theory & Practice**, v. 15, n. 1, p. 85-98, Feb. 2015.

PARSHALL, Cynthia G. et al. **Practical considerations in computer-based testing.** New York: Springer, 2002.

PARTCHEV, Ivailo. **Irtoys:** A Collection of Functions Related to Item Response Theory (IRT), 2016. Disponível em: https://cran.r-project.org/package=irtoys. Acesso em: 24 set. 2018.

PASQUALI, Luiz; PRIMI, Ricardo. Fundamentos da Teoria da Resposta ao Item – TRI. **Avaliação Psicológica**, v. 2, n. 2, p. 99-110, fev. 2003.

PASQUALI, Luiz. **Teoria da Resposta ao Item – TRI**: teoria, procedimentos e aplicações. Brasília: Laboratório de Pesquisa em Avaliação e Medida (LabPAM/UnB), 2007.

PASQUALI, Luiz. **Psicometria**: teoria dos testes na psicologia e na educação. 5. ed. Petrópolis, RJ: Vozes, 2013.

PERRENOUD, Philippe. **Pedagogia diferenciada:** das intenções à ação. Tradução de Patrícia Chittoni Ramos. Porto Alegre: Artmed, 2000.

PITON GONÇALVES, Jean. **A integração de testes adaptativos informatizados e ambientes de tarefas para o aprendizado do inglês instrumental.** 142 p. Dissertação (Mestrado em Ciências da Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2004.

PITON GONÇALVES, Jean. **Desafios e perspectivas da implementação computacional de testes adaptativos multidimensionais para avaliações educacionais.** 177 p. Tese (Doutorado em Ciências da Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2012.

PITON-GONÇALVES, Jean; ALUÍSIO, Sandra Maria. Teste Adaptativo Computadorizado Multidimensional com propósitos educacionais: princípios e métodos. **Ensaio: Avaliação e Políticas Públicas em Educação**, Rio de Janeiro, v. 23, n. 87, p. 389-414, abr./jun. 2015.

PITON-GONÇALVES, Jean; MONZÓN, Andrea Jessica Borges; ALUÍSIO, Sandra Maria. **Métodos de avaliação informatizada que tratam o conhecimento parcial do aluno e geram provas individualizadas.** Trabalho apresentado no XX Simpósio Brasileiro de Informática na Educação, 17 a 20 de novembro de 2009.

PLAJNER, Martin. **Probabilistic models for computerized adaptive testing: study for dissertation thesis.** Prague: Faculty of Nuclear Sciences and Physical Engineering: Czech Technical University in Prague, 2016.

POMMERICH; Mary; SEGALL, Daniel O.; MORENO, Kathleen E. **The nine lives of CAT-ASVAB:** innovations and revelations. Paper presented at Graduate Management Admission Council (GMAC) Conference on Computerized Adaptive Testing, June 2, 2009. Disponível em: http://publicdocs.iacat.org/cat2010/cat09pommerich.pdf. Acesso em: 26 abr. 2015.

PONSODA GIL, Vicente et al. Los tests adaptativos informatizados: investigación actual. **Metodología de las Ciencias del Comportamiento**, suplemento 2004, p. 505-510, 2004.

RECKASE, Mark. Commentary on chapters 5-7: moving from art to science. In: DRASGOW, Fritz (Ed.). **Technology and testing:** improving educational and psychological measurement. New York: Routledge, 2015. (NCME Applications of Educational Measurement and Assessment). p. 174-178.

REIF, Manuel. **PP: Estimation of person parameters for the 1,2,3,4-PL model and the GPCM**. 2014. Disponível em: <https://github.com/manuelreif/PP>.Acesso em 14 dez. 2018.

RENOM, Jordi et al. Investigación en tests adaptativos informatizados. In: OLEA, Julio; PONSODA, Vicente; PRIETO, Gerardo (Ed.). **Tests informatizados:** fundamentos y aplicaciones. Madrid: Pirámide, 1999. (Psicología).

RENOM, Jordi; DOVAL, Eduardo. Tests adaptativos informatizados: estructura y desarrollo. In: OLEA, Julio; PONSODA, Vicente; PRIETO, Geraldo (Ed.). **Tests informatizados:** fundamentos y aplicaciones. Madrid: Pirámide, 1999. p. 127-161.

RENOM, Jordi; DOVAL, Eduardo; SELLÉS, Miguel. Optimización de los TAI mediante el procedimiento de autoarranque. **Revista Electrónica de Investigación y Evaluación Educativa**, v. 4, n. 2-3, p. 1-7, 1998.

REVUELTA, Javier; PONSODA, Vicente. A comparison of item exposure control methods in computerized adaptive testing. **Journal of Educational Measurement**, v. 35, n. 4. p. 311-327, Winter 1998.

REVUELTA, Javier; PONSODA, Vicente; OLEA, Julio. Métodos para el control de las tasas de exposición en tests adaptativos informatizados. **Revista Electrónica de Investigación y Evaluación Educativa**, v. 4, n. 2-2, p. 1-8, 1998.

RHOADES, Kathleen; MADAUS, George. **Errors in standardized tests: a systematic problem**. Chestnut Hill, MA: The National Board on Educational Testing and Public Policy, May 2003.

RIBEIRO, Rui Manuel Bártolo. **Os tempos de latência nas respostas aos itens de testes informatizados:** contributos para a compreensão do processamento cognitivo. 172 f. Dissertação (Mestrado em Comportamento Organizacional) – Instituto Superior de Psicologia Aplicada, Universidade do Minho, Braga, 2001.

RICARTE, Thales Akira Matsumoto. **Teste adaptativo computadorizado nas avaliações educacionais e psicológicas.** 67 f. Dissertação (Mestrado em Ciências da Computação e Matemática Computacional) – Instituto de Ciências Matemática e de Computação, Universidade de São Paulo, São Carlos, 2013.

RIZOPOULOS, Dimitris. ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses, **Journal of Statistical Software**, v.17, n. 5, p. 1-25, nov. 2006. Disponível em: http://www.jstatsoft.org/v17/i05/. Acesso em: 15 out. 2016.

RULISON, Kelly L.; LOKEN, Eric. I've fallen and I can't get up: Can high ability students recover from early mistakes in CAT? **Applied Psychological Measurement,** v. 33, n. 2, p. 83-101, March 2009.

RUSSELL, Michael K. Technology-aided formative assessment of learning. In: ANDRADE, Heidi L.; CIZEK, Gregory J. (Ed.). **Handbook of formative assessment**. Oxon: Routledge, 2010. p. 125-138.

SANTOS, Fabrícia Damando; GUEDES, Leonardo Guerra de Rezende. Testes adaptativos informatizados baseados em Teoria da Resposta ao Item utilizados em ambientes virtuais de aprendizagem. **Novas Tecnologias na Educação**, v. 3, n. 2, p. 1-8, nov. 2005.

SANTOS, Jucelio Soares dos. **Mensuração de habilidades cognitivas preditoras do desenvolvimento de leitura em crianças através de jogos educacionais para dispositivos móveis**. 121 f. Dissertação (Mestrado em Ciência da Computação) – Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande, Campina Grande, 2017.

SÃO PAULO (Município). Secretaria Municipal da Educação. Núcleo Técnico de Avaliação (NTA). **Apresentação do protótipo da plataforma para administração computadorizada da Provinha Brasil – Leitura**: Produto 2 do PRODOC 914 BRZ 1147. Responsável Técnico: Douglas De Rizzo Meneghetti. São Paulo, 2016a.

SÃO PAULO (Município). Secretaria Municipal da Educação. Núcleo Técnico de Avaliação (NTA). **Apresentação da plataforma e o relatório final sobre o uso [...]**: Produto 4 do PRODOC 914 BRZ 1147. Responsável Técnico: Douglas De Rizzo Meneghetti. São Paulo, 2016b.

SÃO PAULO (Município). Secretaria Municipal da Educação. Núcleo Técnico de Avaliação (NTA). **Realizar teste ou simulação do algoritmo para a plataforma [...]**: Produto 3 do PRODOC 914 BRZ 1147. Responsável Técnico: Rodrigo Travitzki Teixeira de Oliveira. São Paulo, 2016c.

SÃO PAULO (município). Secretaria Municipal da Educação. Núcleo Técnico de Avaliação (NTA). **Apresentar o algoritmo e o relatório final, sobre o uso do mesmo na plataforma nas 15 EMEF [...]**: Produto 4 do PRODOC 914 BRZ 1147. Responsável Técnico: Rodrigo Travitzki Teixeira de Oliveira. São Paulo, 2016d.

SASSI, Gilberto Pereira. **Teoria e prática de um Teste Adaptativo Informatizado**. 76 f. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2012.

SASSI, Gilberto Pereira; CÚRI, Mariana. Algoritmos de seleção de itens em testes adaptativos informatizados. In: 19º Simpósio Nacional de Probabilidade e Estatística (Sinape), 2010. **Resumos**. São Pedro: Organização Brasileira de Estatística, 2010, [s.p.].

SAWAKI, Yasuyo. Comparability of conventional and computarized tests of reading in a second language. **Language Learning & Technology**, v. 5, n. 2, p. 38-59, May 2001.

SCACABAROZI, Fernanda Nanci; DINIZ, Carlos Alberto Ribeiro. Uma comparação entre intervalos de credibilidade e o intervalo de confiança clássico para o parâmetro da Distribuição de Poisson.In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 19º, 2010, São Paulo. **Anais eletrônicos**. São Paulo: Associação Brasileira de Estatística, 2010. Disponível em: <http://www.ime.unicamp.br/sinape/19sinape/node/688>. Acesso em: 17 abr. 2019.

SCALISE, Kathleen. **Computer-Based Assessment**: "Intermediate Constraint" Questions and Tasks for Technology Platforms. [S.l.]: University of Oregon, June

2009. Disponível em: <http://pages.uoregon.edu/kscalise/taxonomy/taxonomy.html>. Acesso em: 06 maio 2013.

SCALISE, Kathleen; GIFFORD, Bernard. Computer-Based rAssessment in E-Learning: A Framework for Constructing “Intermediate Constraint” Questions and Tasks for Technology Platforms. **The Journal of Technology, Learning and Assessment**, v. 4, n. 6, June 2006.

SHERMIS, Mark D.; BURSTEIN, Jill (Ed.). **Handbook of automated essay evaluation:** current applications and new directions. Oxon: Routledge, 2013.

SHIN, Chingwei David; CHIEN, Yuehmei; WAY, Walter Denny. **A comparison of three content balancing methods for fixed and variable length computerized adaptive tests**. Vancouver: The National Council on Measurement in Education, April 2012.

SIERRA-MATAMOROS, Fabio Alexánder et al. **Tests adaptativos informatizados. Avances en Medición,** n. 5, p. 157-162, 2007.

SILVA, Vanessa Rufino da. **Avaliação da proficiência em inglês acadêmico através de um teste adaptativo informatizado**. 50 f. Dissertação (Mestrado em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2015.

SIRECI, Stephen G.; ZENISKY, April L. Innovative item formats in computer-based testing: in pursuit of improved construct representation. In: DOWNING, Steven M.; HALADYNA, Thomas M. (Ed.). **Handbook of test development.** Mahwah, NJ: Lawrence Erlbaum, 2006. p. 329-347.

SIRECI, Stephen G.; ZENISKY, April L. Computerized innovative item formats: achievement and credentialing. In: LANE, Suzanne; RAYMOND, Mark R.; HALADYNA, Mark R. (Ed.). **Handbook of test development**. 2nd ed. New York: Routledge, 2016. p. 313-334.

SISTO, Fermino Fendandes. O funcionamento diferencial dos itens. **Psico-USF**, v. 11, n. 1, p. 35-43, jun. 2006.

SOARES, Magda. **Alfabetização:** a questão dos métodos. São Paulo: Contexto, 2016.

SPENASSATO, Débora et al. Vantagens do uso de testes adaptativos computadorizados para avaliação da maturidade do sistema de gestão ambiental de indústrias. **Interciencia**, v.40, n.9, p. 596-603, sept. 2015.

SPRAY, Judith A.; RECKASE, Mark D. **The selection of test items for decision making with a computer adaptive test**. New Orleans: The National Council on Measurement in Education, April 1994.

SPRAY, Judith A.; RECKASE, Mark D. Comparison of SPRT and Sequential Bayes procedures for classifying examinees into two categories using a Computerized Test. **Journal of Educational and Behavioral Statistics**, v. 21, n. 4, p. 405-414, Winter 1996.

STECHER, Brian M. Consequences of large-scale, high-stakes testing on school and classroom practice. In: HAMILTON, Laura; STECHER, Brian M.; KLEIN, Stephen P. (Ed.). **Making sense of test-based accountability in education**. Santa Monica, CA: Rand, 2002. p. 79-100.

STEINBERG, Lynne; THISSEN, David; WAINER, Howard. Validity. In: WAINER, Howard et al. **Computerized adaptive testing:** a primer. 2nd ed. New York: Routledge, 2014. p. 185-230. [Original 2000]

STIGGINS, Richard. Correcting "errors of measurement" that sabotage student learning. In: DWYER, Carol Anne (Ed.). **The future of assessment**: shaping teaching and learning. New York: Lawrence Erlbaum, 2008. p. 229-243.

STOCKING, Martha L. Revising Item Response in Computerized Adaptive Testing: a comparison of three models. **Applied Psychological Measurement**, v. 21, n. 2, p. 129-142, June 1997.

STOCKING, Martha L.; LEWIS, Charles. Controlling item exposure conditional on ability in Computerized Adaptive Testing. **Journal of Educational and Behavioral Statistics**, v. 23, n. 1, p. 57-75, Spring 1998.

STONE, Elizabeth; DAVEY, Tim. **Computer-adaptive testing for students with disabilities:** a review of the literature. Princeton, NJ: Educational Testing Service, Aug. 2011. (Research Report ETS RR-11-32).

STONE, Elizabeth; LAITUSIS, Cara C.; COOK, Linda L. Increasing the accessibility of assessments through technology. In: DRASGOW, Fritz (Ed.). **Technology and testing: improving educational and psychological measurement**. New York: Routledge, 2015. (NCME Applications of Educational Measurement and Assessment). p. 217-234.

STREET, Brian V. Alfabetización y cultura. **Boletín Proyecto Principal de Educación en América Latina y el Caribe**, Santiago, n. 32, p. 39-46, Dic. 1993.

THISSEN, David. Reliability and measurement precision. In: WAINER, Howard et al. **Computerized adaptive testing:** a primer. 2nd ed. New York: Routledge, 2014. p. 159-184. [Original 2000]

THOMPSON, Nathan A. Item Selection in Computerized Classification Testing. **Educational and Psychological Measurement**, v. 69, n. 5, p. 778-793, Oct 2009.

THOMPSON, Nathan A. **Adaptive testing:** is it right for me? Saint Paul, MN: Assessment Systems Corporation, 2010.

THOMPSON, Nathan A.; WEISS, David J. A framework for the development of computerized adaptive tests. **Practical Assessment, Research & Evaluation**, v. 16, n. 1, Jan. 2011.

TOKARNIA, Mariana. **Provinha Brasil terá apenas versão digital por restrições financeiras, diz Inep**. Agência Brasil, Brasília, 08 ago. 2016. Disponível em: <http://agenciabrasil.ebc.com.br/educacao/noticia/2016-08/provinha-brasil-tera-apenas-versao-digital-por-restricoes-financeiras-diz> Acesso em: 01 out. 2018.

TRAVITZKI, Rodrigo et al. How to build a Computerized Adaptive Test with free software and pedagogical relevance? In: **PROCEEDINGS of IAC 2018 in Vienna**: Teaching, Learning and E-learning (IAC-TLEI 2018). Prague: Czech Institute of Academic Education z.s., 2018. p. 117-126.

UBRIACO, F. E. de C. A. Interpretação de escalas de proficiência com utilização do método do marcador. **Estudos em Avaliação Educacional**, São Paulo, v. 23, n. 52, p. 86-105, maio/ago. 2012.

UNIVERSAL. **Design for Computer-Based Testing (UD-CBT) Guidelines.** Upper Saddle River, NJ: Pearson; Peabody, MA: CAST, Oct. 2010.

UNIVERSIDADE DE SÃO PAULO. Sistema Integrado de Bibliotecas da USP. **Diretrizes para apresentação de dissertações e teses da USP**: parte I (ABNT)/Sistema Integrado de Bibliotecas da USP, Vânia Martins Bueno de Oliveira Funaro, (Coord.) Vânia Martins Bueno de Oliveira Funaro et al. 3. ed. rev. ampl. mod. São Paulo: SIBiUSP, 2016. 100 p. (Cadernos de Estudos, 9).

URRY, Vern W. Tailored testing: a successful application of latent trait theory. **Journal of Educational Measurement**, v. 14, n. 2, p. 181-196, Summer 1977.

VAN DER LINDEN, W.J.; GLAS, Cees A. W. (Ed.). **Computerized adaptive testing**: theory and practice. Dordercht: Kluwer Academic Publishers, 2010a.

VAN DER LINDEN, W.J.; GLAS, Cees A. W. (Ed.). **Elements of adaptive testing**. New York: Springer, 2010b. (Statistics for Social and Behavioral Sciences).

VAN DER LINDEN, Wim J. Some alternatives to Sympson-Hetter item-exposure control in Computerized Adaptive Testing. **Journal of Educational and Behavioral Statistics**, v. 28, n. 3, p. 249-265, Fall 2003.

VAN DER LINDEN, Wim J.; VELDKAMP, Bernard P. Constraining item exposure in Computerized Adaptive Testing with shadow tests. **Journal of Educational and Behavioral Statistics**, v. 29, n. 3, p. 273-291, Fall 2004.

VAN DER LINDEN, Wim J. Optimal test assembly. In: LANE, Suzanne; RAYMOND, Mark R.; HALADYNA, Mark R. (Ed.). **Handbook of test development**. 2nd ed. New York: Routledge, 2016. p. 507-530.

VEERKAMP, Wim J. J.; BERGER, Martijn P. F. Some new item selection criteria for adaptive testing. **Journal of Educational and Behavioral Statistics**, v. 22, n. 2, p. 203-22, Summer 1997.

VELDKAMP, Bernard P.; MATTEUCCI, Mariagliulia. Bayesian Computerized Adaptive Testing. **Ensaio: Avaliação e Políticas Públicas em Educação**, Rio de Janeiro, v. 21, n. 78, p. 57-82, jan./mar. 2013.

VELDKAMP, Bernard P.; VAN DER LINDEN, Wim J. Implementing Sympson-Hetter item-exposure control in a shadow-test approach to Constrained Adaptive Testing. **International Journal of Testing**, n. 8, p. 272-289, 2008.

VIEIRA JUNIOR, Niltom. **Planejamento de um ambiente virtual de aprendizagem baseado em interfaces dinâmicas e uma aplicação ao estudo de potência elétrica**. 232 p. Tese (Doutorado em Engenharia elétrica) – Faculdade de Engenharia de Ilha Solteira, Universidade Estadual Paulista, Ilha Solteira, 2012.

VIANNA, Heraldo Marelim. **Testes em educação**. São Paulo: Ibrasa, 1973.

VIANNA, Heraldo Marelim. Avaliações nacionais em larga escala: análises e propostas. **Estudos em Avaliação Educacional**, São Paulo, n. 27, p. 41-76, jan./jun. 2003.

WAINER, Howard et al. **Computerized adaptive testing**: a primer. 2nd ed. Mahwah, NJ: Lawrence Erlbaum, 2000a.

WAINER, Howard. CATs: whither and whence. **Psicológica**, v. 21, n. 1, p. 121-133, 2000b.

WAINER, Howard. Introduction and history. In: WAINER, Howard et al. **Computerized adaptive testing:** a primer. 2nd ed. New York: Routledge, 2014. p. 1-21. [Original 2000]

WANG, Hong; SHIN, Chingwei David. Comparability of computerized adaptive and paper-pencil tests. **Test, Measurement & Research Services Bulletin**, Issue 13, p. 1-7, March 2010.

WANG, Shiyu et al. Hybrid computerized adaptive testing: from group sequential design to fully sequential design. **Journal of Educational Measurement**, v. 53, n. 1, p. 45-62, Spring 2016.

WANG, Tianyou; KOLEN, Michael J. Evaluating Comparability in Computerized Adaptive Testing: Issues, Criteria and an Example. **Journal of Educational Measurement**, v. 38, n. 1, p. 19-49, Spring 2001.

WARM, Thomas A. Weighted likelihood estimation of ability in item response theory. **Psychometrika**, v. 54, n. 3, p. 427-450, Sept. 1989.

WAY, Walter D. et al. From standardization to personalization: the comparability of scores based on different testing conditions, modes and devices. In: DRASGOW, Fritz (Ed.). **Technology and testing:** improving educational and psychological measurement. New York: Routledge, 2015. (NCME Applications of Educational Measurement and Assessment). p. 260-284

WEISS, David J. Improving measurement quality and efficiency with adaptive testing. **Applied Psychological Measurement**, v. 6, n. 4, p. 473-492, Fall 1982.

WEISS, D. J. Better data from better measurements using computerized adaptive testing. **Journal of Methods and Measurement in the Social Sciences**, v. 2, n. 1, p. 1-23, 2011.

WEISS, David J.; KINGSBURY, G. Gage. Application of Computerized Adaptive Testing to educational problems. **Journal of Educational Measurement**, v. 21, n. 4, p. 361-375, Winter 1984.

WILLIAMS, Raymond. **Cultura**. 2. ed. Tradução de Lólio Lourenço de Oliveira, Rio de Janeiro: Paz e Terra, 2000.

WILLIAMSON, David M.; MISLEVY, Robert J.; BEJAR, Isaac I. (Ed.). **Automated scoring of complex tasks in computer-based testing**. Mahwah, NJ: Lawrence Erlbaum, 2006.

WISE, Steven L.; KINGSBURY, G. Gage. Practical issues in developing and maintaining a computerized adaptive testing program. **Psicológica**, n. 21, p. 135-155, 2000.

YAN, Duanli; LEWIS, Charles; VON DAVIER, Alina A. Overview of computarized multistage tests. In: YAN, D.; VON DAVIER, A. A.; LEWIS, C. (Ed.). **Computerized multistage testing: theory and applications.** Boca Raton, FL: CRC Press, 2014. (Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences Series). p. 3-20.

YAN, Duanli; VON DAVIER, Alina A.; LEWIS, Charles (Ed.). **Computerized multistage testing: theory and applications**. Boca Raton, FL: CRC Press, 2014. (Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences Series).

ZHANG, Susu; CHANG, Hua-Hua. From smart testing to smart learning: how testing technology can assist the new generation of education. **International Journal of Technology and Learning**, v. 1, n. 1, p. 67-92, 2016.

ZHU, Weimo. Science and art of setting performance standards and cutoff scores in Kinesiology. **Research Quarterly for Exercise and Spor**t, v. 84, n. 4, p. 456-468, 2013.

ANEXO A – Matriz de Referência para Avaliação da Alfabetização e do Letramento Inicial

Eixo	Habilidade/Descritor	Especificidade da habilidade (níveis de complexidade)
Apropriação do sistema de escrita	D1 – reconhecer letras.	D1.1 – diferenciar letras de outros sinais gráficos.
		D1.2 – identificar as letras do alfabeto.
		D1.3 – identificar diferentes tipos de letras.
	D2 – reconhecer sílabas.	D2.1 – identificar número de sílabas a partir de imagens.
	D3 – estabelecer relação entre unidades sonoras e suas representações gráficas.	D3.1 – identificar vogais nasalizadas.
		D3.2 – identificar relação entre grafema e fonema (letra/som – com correspondência sonora única; ex.: p,b,t,d,f).
		D3.3 – identificar relação entre grafema e fonema (letra/som – com mais de uma correspondência sonora; ex.: c e g).
		D3.4 – reconhecer, a partir da palavra ouvida, o valor sonoro de uma sílaba.
		D3.5 – reconhecer, a partir de imagem, o valor sonoro de uma sílaba.
Eixo	**Habilidade/Descritor**	**Especificidade da habilidade (níveis de complexidade)**
Leitura	D4 – ler palavras.	D4.1 – estabelecer relação entre significante e significado.
	D5 – ler frases.	D5.1 – ler frases.
	D6 – localizar informação explícita em textos.	D6.1 – localizar informação explícita em textos.
	D7 – reconhecer assunto de um texto.	D7.1 – Reconhecer o assunto do texto com apoio das características gráficas e do suporte.
		D7.2 – Reconhecer o assunto do texto com base no título.
		D7.3 – Reconhecer o assunto do texto a partir da leitura individual (sem apoio das características gráficas ou do suporte).
	D8 – identificar a finalidade do texto.	D8.1 – Reconhecer a finalidade do texto com apoio das características gráficas do suporte ou do gênero.
		D8.2 – Reconhecer a finalidade do texto a partir da leitura individual (sem apoio das características gráficas do suporte ou do gênero).
	D9 – estabelecer relação entre partes do texto.	D9.1 – Identificar repetições e substituições que contribuem para a coerência e coesão textual.
	D10 – inferir informação.	D10.1 – Inferir informação.

Fonte: (BRASIL, 2016a)

ANEXO B – Quadros descritivos dos itens do BI do TAI da PB – Leitura

Quadro 6 – Descrição dos itens do TAI da PB – Leitura, teste 1, edição de 2015, por eixo de habilidade, descritor e posição no teste convencional

Eixo da Matriz	Código do descritor	Descritor	Posição do item no teste impresso
1	D3.2	Identificar relação entre grafema e fonema (letra/som - com correspondência sonora única; ex.: p,b, t, d, f).	1
2	D4.1	Estabelecer relação entre significante e significado.	2
1	D1.1	Diferenciar letras de outros sinais gráficos.	3
1	D1.2	Identificar as letras do alfabeto.	4
1	D3.3	Identificar relação entre grafema e fonema (letra/som - com mais de uma correspondência sonora; ex.: "c" e "g").	5
1	D3.4	Reconhecer, a partir de palavra ouvida, o valor sonoro de uma sílaba.	6
2	D5.1	Ler Frases.	7
1	D1.3	Identificar diferentes tipos de letras.	8
1	D3.5	Reconhecer, a partir de imagem, o valor sonoro de uma sílaba.	9
2	D6.1	Localizar informação explícita em textos.	10
1	D2.1	Identificar número de sílabas a partir de imagens.	11
1	D3.5	Reconhecer, a partir de imagem, o valor sonoro de uma sílaba.	12
2	D7.1	Reconhecer o assunto do texto com apoio das características gráficas e do suporte.	13
2	D7.2	Reconhecer o assunto do texto com base no título.	14
2	D8.1	Reconhecer a finalidade do texto com apoio das características gráficas do suporte ou do gênero.	15
2	D8.1	Reconhecer a finalidade do texto com apoio das características gráficas do suporte ou do gênero.	16
2	D9.1	Identificar repetições e substituições que contribuem para a coerência e coesão textual.	17
2	D7.1	Reconhecer o assunto do texto com apoio das características gráficas e do suporte.	18
2	D7.3	Reconhecer o assunto do texto a partir da leitura individual (sem apoio das características gráficas ou do suporte).	19
2	D10.1	Inferir informação.	20

Quadro 7 – Descrição dos itens do TAI da PB – Leitura, teste 2, edição de 2015, por eixo de habilidade, descritor e posição no teste convencional

Eixo da Matriz	**Código do descritor**	**Descritor**	Posição do item no teste impresso
2	D4.1	Estabelecer relação entre significante e significado.	1
1	D3.2	Identificar relação entre grafema e fonema (letra/som - com correspondência sonora única; ex.: p,b, t, d, f).	2
1	D1.1	Diferenciar letras de outros sinais gráficos.	3
1	D3.3	Identificar relação entre grafema e fonema (letra/som - com mais de uma correspondência sonora; ex.: "c" e "g").	4
2	D5.1	Ler Frases.	5
1	D1.3	Identificar diferentes tipos de letras.	6
1	D2.1	Identificar número de sílabas a partir de imagens.	7
1	D3.4	Reconhecer, a partir de palavra ouvida, o valor sonoro de uma sílaba.	8
2	D6.1	Localizar informação explícita em textos.	9
2	D7.1	Reconhecer o assunto do texto com apoio das características gráficas e do suporte.	10
2	D7.2	Reconhecer o assunto do texto com base no título.	11
2	D7.3	Reconhecer o assunto do texto a partir da leitura individual (sem apoio das características gráficas ou do suporte).	12
2	D8.1	Reconhecer a finalidade do texto com apoio das características gráficas do suporte ou do gênero.	13
1	D3.5	Reconhecer, a partir de imagem, o valor sonoro de uma sílaba.	14
2	D10.1	Inferir informação.	15
2	D10.1	Inferir informação.	16
2	D7.3	Reconhecer o assunto do texto a partir da leitura individual (sem apoio das características gráficas ou do suporte).	17
2	D7.3	Reconhecer o assunto do texto a partir da leitura individual (sem apoio das características gráficas ou do suporte).	18
2	D8.2	Reconhecer a finalidade do texto a partir da leitura individual (sem apoio das características gráficas do texto-base ou do gênero).	19
2	D9.1	Identificar repetições e substituições que contribuem para a coerência e coesão textual.	20

ANEXO C – Resultado por escola no TAI da PB – Leitura

Escola	1º ano			2º ano			Total
	média	desvioP	Total	média	desvioP	Total	
Angola	434,15	71,16	87	494,25	81,69	84	171
África do sul	408,06	55,71	71	471,04	72,32	57	128
Moçambique	407,17	66,89	62	489,57	96,02	128	190
Libéria	417,23	61,71	68	482,30	79,10	70	138
Argélia	416,79	60,58	45	486,95	70,69	48	93
Costa do Marfim	411,04	42,97	70	483,21	78,87	80	150
Libia	393,22	59,81	29	501,18	79,18	40	69
Cabo Verde	421,86	63,06	76	500,54	67,04	78	154
República do Congo	417,79	63,69	45	506,03	86,96	103	148
Camarões	438,99	73,30	52	530,10	67,45	90	142
Etiópia	329,24	-	1	518,16	68,24	76	77
Benim	412,47	64,97	49	486,80	71,04	52	101
Ruanda	416,28	49,85	82	486,13	74,13	83	165
Egito	412,68	60,04	49	480,76	71,04	92	141
Marrocos	410,09	54,70	37	500,22	81,77	79	116
Total	**-**	**-**	**823**	**-**	**-**	**1160**	**1983**

ANEXO D – Registro das aplicações em papel e lápis e eletrônica da PB – Leitura

1. Emef Angola

Considerações gerais	Considerações Específicas		
Cada equipe da gestão da Emef, durante a aplicação, estava presente.	Eletrônica	Turma 1	Alunos quietos (comportados). Alguns apresentaram dificuldades nas questões de leitura. Dois alunos não compreenderam a pronúncia nas duas vezes e acabaram chutando as respostas.
		Turma 2	Estavam agitados, ansiosos e curiosos, tentando ver como funcionava. Quando a professora começou a explicar ficaram quietos, tentaram responder às questões no *tablet* do outro, leram muito alto, falando a resposta a partir da questão 11 ficaram silenciosos e com pressa de acabar. A primeira aluna acabou às 14:47 (obs.: a prova começou às 14:39).
	Papel	2ºA	Prof. não recebeu o guia do Inep, embora conhecesse a aplicação da provinha. O responsável pelo registro forneceu o guia Inep. Crianças atentas. Duas crianças começaram a mostrar dispersão por volta da Q 12.
		2ºB	Antes da aplicação os alunos estavam tranquilos, sentados separadamente e prestando atenção nas orientações. A aplicação foi bem direcionada e os alunos não mostraram dúvidas. Quando chegou na questão 17 o efeito canseira se manifestou em vários alunos que passaram a reclamar das questões.
		2ºC	A professora inicialmente realizava a leitura de toda a questão com os alunos. Fiz uma intervenção de como deveria ler as questões segundo o guia de aplicação.

2. Emef África do Sul

Considerações gerais	Considerações Específicas		
Um aluno expôs que não sabia ler, quando começaram as questões que envolviam leitura de textos. Foi necessário orientar para que respondesse as questões com alternativas que considerava corretas. O tempo foi restrito para uma aluna que começou na segunda etapa.	Eletrônica	Turma A	Alunos tranquilos, bem comportados.
		Turma B	Crianças agitadas. Primeira criança: 17:03.
	Papel	2ºA	Sala inquieta. Uma das alunas falou em voz alta: “Professora, eu preciso de ajuda, não sei ler” e falou isso várias vezes. Ela acabava olhando a da colega de trás e colocava a resposta igual. Questão 7: uma das alunas falou a resposta em voz alta. Questão 8: aluna apontou resposta e a professora balançou a cabeça. Questão 9: foi repetida mais de uma vez e uma das crianças chorou (chorou em quase todas).
		2º C	--

3. Emef Moçambique

Considerações gerais	Considerações Específicas		
Professoras que estão se aposentando.	Eletrônica	Turma 1	Quando a conversa estava muito alta, os alunos disseram que não conseguiram ouvir a fala do *tablet*.
	Papel	2ºA	Foi realizada uma reunião de orientação com os docentes aplicadores. Os alunos apresentaram dificuldades a partir do item 8. A partir do 10º item os alunos começaram a se dispersar. A partir do item 8 a sala começou a ter dificuldade.
		2ºB	Erro na questão 14: a professora leu o texto da questão. Após erro, a professora foi orientada para aplicar corretamente os seguintes itens. Ao final do prazo, professora pediu para que verificassem se haviam respondido todos os itens. Alunos agitados antes e durante a aplicação. Reações oriundas do nível de proficiência observadas, por exemplo: alunos que liam melhor estavam à frente, enquanto colegas se detiam na que estava sendo respondida.
		2º C	Presença de estagiária de educação especial B. Crianças sentadas em suas carteiras. Às 9h:44 um aluno foi repreendido. Professora explicou que sílabas são os pedacinhos. Às 10:10 professora começou a recolher as provas. Um aluno não tinha colocado o nome.

4. Emef Argélia

Considerações gerais	Considerações Específicas		
Nesta EMEF, alguns já tinham respondido ao TAI e, por isto, hoje farão a prova eletrônica.	Eletrônica	Turma 1	14 alunos participaram da prova. Alunos agitados, houve problema com a impressão dos RA dos alunos pois a escola imprimiu o EOL. Aluno M. não conseguiu acessar a prova com seu RA, então foi utilizado o RA 5706000. Os alunos que iam terminando a prova permaneceram na sala. Muitos alunos com dificuldade para realizar a prova. A professora L, permaneceu conosco durante a aplicação.
		Turma 2 2º A	Aluno não sabe ler, teve auxílio da professora com a leitura; ele terminou a prova às 14:45, foi o primeiro (obs.: a prova começou às 14:33. A prova demorou a começar, pois a lista com RA estava errada. A prova teve um ótimo desenvolvimento.
	Papel	2ºA	A aplicação da prova ocorreu sem intercorrências. Obs.: a sala não tem aluno com deficiência.
		2ºB	Primeiras questões 'muito fácil' até 08. A aluna Vitoria apresentou comportamento não condizente com situação de prova. A professora diversas vezes parava a aplicação para chamar sua atenção.
		2ºC	--

5. Emef Camarões

Considerações gerais	Considerações Específicas		
Uma aluna não sabia o que significava iniciar. Dois alunos não entenderam a questão da letra B/T. Um aluno não escutou a frase que tinha que ser identificada.	Eletrônica	Turma 1	Primeiro aluno: 15h14min; último: 15h24min.
		Turma 2	Prova tranquila com 7 alunos. Crianças quietas. Durante a aplicação ficaram em silêncio. Não tiveram dúvidas e nem dificuldades.
		Turma 3	Aluna G. com deficiência intelectual comprovada. Dividimos o grupo / 10 alunos fizeram na segunda aplicação; correu tudo bem.
		Turma 4	Alunos agitados, houve dificuldade para conectar os *tablets*, aluno teve dúvida sobre o que seria (inicial). Alguns alunos tiveram dúvidas quanto à leitura do texto (referente aonde clicar) houve a necessidade de dividir a turma para aplicar a prova pois a internet não conectava. Também houve a necessidade de ditar os números de EOL pois vieram em lista, não em tiras.
	Papel	2ºA	A professora informou corretamente as orientações. Mesmo a professora não tendo lido o guia, ela já havia aplicado e já tinha conhecimentos. Leu a questão 11 para os alunos. Houve troca de professor na sala (saiu Mariane e entrou a Tatiana). A professora Tati conhecia a prova, já aplicou mais eficiência. Questão 16: aparentavam cansaço. Responder sem ler: 80% da sala/ antes de a professora ler o enunciado.
		2ºB	Sem registro
		2ºC	-

6. Emef Líbia

Considerações gerais	Considerações Específicas		
2ºs anos com período integral.	Eletrônica	Turma 1	As crianças foram pacientes na aplicação. O primeiro aluno acabou às 16h48min e o último aluno às 17:00 (a prova iniciou às 16:40).
		Turma 2	-
	Papel	2ºA	Algumas crianças começaram a ter dificuldades a partir da questão 11. Uma aluna ficou muito ansiosa por não conseguir responder e chorou. A professora tentou consolá-la.
		2ºB	Um dos alunos não esperava a professora, ia respondendo antes e indo para as próximas questões. A partir da questão 9 alunos começaram a ter dificuldade. Questão 11: professora leu o exto. Professora teve que explicar o que é a palavra 'destacada', pois alguns não sabiam o que era uma palavra destacada. Questão 13: alunos responderam em voz alta a resposta da questão. Às 17:05 um dos alunos terminou a prova (que começou às 16:37)
		2ºC	-

7. Emef Costa do Marfim

Considerações gerais	Considerações Específicas		
Sem registro.	Eletrônica	Turma 1	Turma calma, comportada e organizada. Todos os alunos prestaram atenção nas instruções. Exatamente quando a prova iniciou, fui chamado para auxiliar a ver os fones de ouvido da outra turma, pois estavam com volume muito baixo. Ajude a testar cada *tablet* dos alunos. Voltei para a turma inicial faltando cerca de 5 min para o término. Obs.: a aluna R. inicia sua prova às 14:58, pois seu *tablet* não estava funcionando.
		Turma 2	Alguns alunos iniciaram antes de testar o som (coordenadora?). Começou instrução sem duas crianças que foram ao banheiro. Um *tablet* travou/ criança não alfabetizada. Coordenadora fez a prova de uma criança.
	Papel	2ºA	Inicialmente os alunos acharam os itens muito fáceis, dispensando durante a leitura pela professora até o item 10. Dúvida da maioria dos alunos na questão 2. Alunos acham difícil os itens que tem que ler silenciosamente (sozinhos): 11, 13, 14, 16, 17, 18, 19. Um aluno terminou a prova antes de todos, sem seguir as instruções da professora. Um aluno com deficiência intelectual não teve nenhum interesse em fazer a prova, ficou brincando com adesivos na própria mesa. Mas no fim da prova, o aluno NEE começou a deitar no chão, mexer nos materiais dos colegas, exigindo mais a atenção da professora.
		2ºB	A professora fez um alongamento/relaxamento para acalmar os alunos antes da prova. Explica o procedimento da prova – seguiu o guia, atentou bem as crianças. Questão 11: realizou a leitura da questão. Os alunos permaneceram calmos, não se levantaram.
		2ºC	14:21: entrega da prova, os alunos colocaram somente os nomes; 14:28: instruções para os alunos dadas pela professora; 14:30: início da prova: alunos calmos. Alguns já folhearam tentando responder antes que a professora lesse as questões. Alunos com bastante tempo para ler os textos com o passar do tempo começaram a ficar inquietos na carteira; aluno desistiu da prova. 14:57: fim da prova.

8. Emef Benim

Considerações gerais	Considerações Específicas		
Nenhum registro.	Eletrônica	Turma 1	Antes da aplicação os alunos estavam agitados com a novidade (*tablet*). Uma aluna ficou muito nervosa e começou a chorar com medo de fazer a prova. Poucas crianças tiveram dúvidas. Todos alunos do 2A. O professor não estava presente após as 14:15; os alunos que já haviam terminado ficaram bem agitados. Para conter os alunos, Renata contou uma história e fez uma brincadeira para que a criança que estava nervosa terminasse a prova.
		Turma 2	-
	Papel	2ºA	-
		2ºB	Não houve ocorrências.
		2ºC	-

9. Emef Etiópia

Considerações gerais	Considerações Específicas		
Aplicação 1: Turmas 2A e 2C foram agrupadas na sala da turma 2A. Prof. da turma 2A só chega às 10h, por isso Profa. do 2C, fará a aplicação. Contudo, ela não tinha recebido o manual de aplicação e precisou de um tempo para lê-lo. Aplicação 2: Prova iniciada 09h59min. Aluna T. realizou a prova com RA do aluno S. Último aluno terminou às 10:29.	Eletrônica	Turma 1 (2A)	Aluna autista se recusou a fazer a prova, ficou brincando de desenhar no *tablet*. Primeira aluna terminou 09h54min. Aluno comentou que ouviu perfeitamente a voz do *tablet*. Seis alunos fizeram com RA de outros. Distribuição errada dos alunos pela escola.
		Turma 2	-
	Papel	2ºA e 2º C	Prof. do 2ª chegou às 09:30. Crianças atentas.
		2ºB	Houve interrupção na questão 5 pois chegaram alguns alunos que não estavam na sala. Após o início da prova da prova para os que chegaram depois, todos seguiram juntos. Às 09:35 a professora foi novamente interrompida para dar informações à funcionária.
		2C	-

10. Emef Cabo Verde

Considerações gerais	Considerações Específicas		
Não teve registro.	Eletrônica	Turma 1	*Tablet* 17 sem síntese de voz. Turma agitada durante as instruções. As questões de texto: um dos alunos não lia, simplesmente colocava qualquer resposta e passava para a próxima.
		Turma 2	O aluno J. H. B. fez prova com o nome (RA) de Maria Clara Santos. Vinte alunos fizeram a prova.
	Papel	2ºA	Os alunos perguntaram o 'por quê' de tantos textos.
		2ºB	Questão 2: dificuldade em entender as letras ditadas D e P. Sala bem quieta/ alunos calmos. Muito barulho exterior no corredor. Questão 13: texto longo. Questão 19: alguns não conheciam a palavra 'pálidos'.
		2ºC	Alunos tranquilos (14 alunos). Professora colocou e explicou a questão-exemplo. Q9: Professora explicou o que é sílaba ('quando se abre a boca para falar').

11. Emef Libéria

Considerações gerais	Considerações Específicas		
Aplicação 1: Dificuldades para definir as salas para aplicação da prova. Aplicação 2: Quando a prova é iniciada em outro tablet (problema de bateria ou conexão), a prova precisa ser reiniciada.	Eletrônica	Turma 1: 2C e parte do 2A.	Antes da aplicação os alunos mostraram-se eufóricos, entusiasmados e sem dificuldades para utilizar os *tablets*. O primeiro aluno terminou a prova às 10:05. A aluna E. interrompeu a prova várias vezes, reiniciou a prova às 10:14 por problemas de conexão, reiniciou a prova em outro *tablet*. Os demais alunos realizaram a prova sem problemas ou dificuldades.
		Turma 2	Problemas com o som baixo. Aluno B. teve que reiniciar a prova em outro *tablet* devido a falta de conexão. Iniciou às 09:40 e terminou às 10:00. Aluno J.. Realizou a prova com o RA da aluna Y. Dúvida: L– autismo – leu?
	Papel	2ºA	A professora organizou a sala para que o ambiente favorecesse a aplicação, conferindo, inclusive, os materiais dos alunos. Perguntas: 'O que é inicial'. A turma começou a conversar a partir do item 11.
		2ºB	A professora da sala faltou, quem aplicou a prova foi a professora de Inglês, que nunca tinha aplicado a Provinha. Questões 2 e 11: professora leu mais de duas vezes. Sala bem agitada (alunos falaram durante toda a prova). Até questão 8: alunos acharam muito fácil. A partir da questão 13 começaram a ficar cansados, reclamando que tinha muito texto. Alunos com dificuldade em saber o que é 'assunto principal' do texto'. Fala de um aluno: 'O assunto principal é a primeira resposta, né?'.
		2ºC	Os alunos muito inquietos. 12 alunos + 1 (com acompanhante) múltiplas deficiências. Professora solicitou que colocassem seus nomes. Professora não avisou que o primeiro item era questão-exemplo. Q2 Professora disse P de Paulo. Porta aberta até Q8, com entrada de barulho externo. Q10: leu 3 x. Q 11: explicou o que significa a substituição, com exemplo, e leu 3x. Enquanto algumas (3) crianças evidenciavam dificuldades, outras queriam adiantar. Professora disse para um aluno que precisava aprender a esperar. Durante a Q19 a dispersão era visível. Pelo menos 5 alunos entregaram a prova logo após a comanda da Q20.

12. Emef Ruanda

Considerações gerais	Considerações Específicas		
Nenhum registro.	Eletrônica	Turma 1	Um aluno começou a prova antes dos demais. Início para os demais: 10:47. Alguns *tablets* caíram no início da prova. Aluna número 7: *tablet* caiu várias vezes.
		Turma 2	Alunos atentos às instruções. Primeiro aluno: 14:18. Total de alunos: 14 (2B). último aluno: 14:39.
		Turma 3	Alunos muito agitados. Muitos *tablets* tiveram dificuldade para carregar desde a primeira pergunta, travando e dois ficaram off-line. O aluno D. teve dificuldade para conectar, sendo necessário considerar a última prova deste aluno.
	Papel	2ºA	Nenhum registro.
		2ºB	Os alunos estavam conversando com a professora muito ansiosos e ela os acalmou durante a prova. Alguns alunos estavam perdidos nas questões enquanto a professora estava na questão 8 eles estavam na 6 ou 7, quando a professora percebeu ela os aguardou até que conseguissem acompanhar todo o grupo. Todas as dúvidas foram solucionadas e os alunos não tiveram dificuldades em nenhum momento.
		2ºC	Dos doze alunos presentes um com deficiência auditiva com acompanhamento do intérprete educacional. Às 10:36 a professora iniciou as instruções das questões lendo-as duas vezes. Os alunos não apresentaram dificuldades quanto ao acompanhamento das questões propostas. A intérprete leu as questões para o aluno com deficiência. Às 11:03 os alunos concluíram a prova, inclusive o que tem deficiência auditiva.
		2º C	Alunos tranquilos, acharam a prova fácil. A professora precisou ser orientada para não ler os textos. O aluno de DA não conseguiu fazer toda a prova, pois foi alfabetizado em libras. Q9: a intérprete não conseguiu explicar, pois sílaba ele não entende e também não conseguiu fazer a questão referente à primeira letra, pelo mesmo motivo.

13. Emef República do Congo

Considerações gerais	Considerações Específicas		
Nenhum registro.	Eletrônica	Turma 2ºD	Alguns alunos reclamaram do som baixo. A primeira aluna terminou às 17:23; o segundo aluno terminou às 17:25. *Tablet* 51 está com Chrome desatualizado.
		Turma 2	O login do aluno M.não entrou e foi substituído por 6533210. Começou às 17h32min e o primeiro aluno terminou às 17h39min. Mais de 50% da sala já tinha saído às 17:48. Uma criança perguntou o que é 'pálido'. Ás 17h51min somente um aluno estava em prova. Às 17h52min terminou o último aluno. Os alunos mostraram gostar de prova no *tablet*. Ficaram muito agitados e 5 crianças apertaram o início antes da orientação e dois iniciaram a prova sem a orientação. A CP Patrícia acompanhou a aplicação. Um *tablet* estava sem google Chrome.
	Papel	2ºA	Questão 8: a professora repetiu mais de duas vezes. Alunos agitados repetiam a resposta em voz alta, fazendo com que, quem não sabia, colocasse o que o colega falava (até a questão 11 apenas). Os alunos tiveram dificuldades com as questões de textos (acabou sendo mais demorada, pois os alunos eram lentos). Faziam as questões que eles sabiam

			antes da orientação da professora e pulavam para a próxima. Um dos alunos terminou a prova às 17:11.
		2ºB	Os alunos estavam tranquilos, mas não queriam fazer a prova em lápis e papel. Contudo, depois da explicação eles fizeram a prova em silêncio. Não apresentaram dúvidas até a questão 14.
		2ºC	-
		2ºD	Os alunos estavam chateados porque também queriam fazer a prova no *tablet*. A professora teve o cuidado de não dar as respostas. A professora deu tempo suficiente para a leitura dos textos, porém, após explicar a pergunta, não esperou muito tempo para pensarem na resposta. Um aluno, Tiago, com dificuldade de aprendizagem (aparentemente o aluno tem alguma deficiência) não conseguia fazer a prova quando iniciaram os textos. Este mesmo aluno estava sentado na última carteira (aluno com deficiência cognitiva). Os alunos tiveram dificuldade com o vocabulário da prova. Os alunos ficaram cansados com tantos textos.

14. Emef Egito

Considerações gerais	**Considerações Específicas**		
Nenhum registro.	Eletrônica	Turma 1 (2A, 2B, 2C)	Do 2ª foram 16 alunos presentes, do 2B foram 15 presentes, 3 faltaram e 2 foram substituídos, do 2C 15 presentes e 1 faltou. A turma era inquieta. O primeiro aluno saiu às 17:05. A partir das questões de textos os alunos ficaram quietos e demoraram um pouco mais.
		Turma 2	-
	Papel	2ºA	13 alunos fizeram a prova. Havia muito barulho no corredor. A professora leu , na Q9, a palavra 'planeta' silabadamente. A professora leu o texto das questões: 9, 11, 13, 14, 16, 18, 19, 20. A exceção de um aluno, um pouco inquieto, os alunos estavam concentrados.
		2ºB	Sala com alunos agitados/falantes. Questões 1, 2, 3, 4: os alunos acharam muito fáceis. Questão 9: acharam muito difícil, a maioria não sabe o que é sílaba. O professor explicou que são os pedaços das palavras. Dificuldade na questão 11 (ela). Questão 13: texto longo e difícil. A partir da questão 14 houve troca de professor, sendo a professora de Educação Física que aplica a Provinha. Os alunos continuam agitados, falando durante a prova. A partir da questão 16 – acharam os textos longos. A professora de Educação Física gritou várias vezes para as crianças ficarem quietas.
		2ºC	Os alunos não gostaram de fazer a prova, questionaram a escolha e ficaram inquietos. Durante a aplicação, eles fizeram quietos e não houve problema. A professora não leu nenhuma palavra com entonação e andou pela sala a todo momento para garantir que todos respondessem. A questão 9 não ficou clara para os alunos. A questão 11 gerou muitas dúvidas quanto ao significado da palavra 'destacada' e quanto ao enunciado.

15. Emef Marrocos

Considerações gerais	Considerações Específicas		
Nenhum registro.	Eletrônica	Turma 1	16 alunos + 8.
		Turma 2	20 alunos. Problema de conectividade múltipla. Alunos questionaram sobre o procedimento após marcarem errado uma questão. Questão 14 respondida aleatoriamente pelo aluno de RA 5800376. Não ouviu a comanda. Por problema de conectividade um aluno não concluiu a prova.
	Papel	2ºA	A professora acompanhou o aluno que apresenta deficiência não identificada. Os alunos apresentaram dificuldades em algumas questões, ex.: 7, 11, 13, 16. A maioria dos alunos, para responder às questões, fez a leitura oral. A professora se dirigia até os alunos para auxiliá-los nas respostas das questões. Antes de avançar para a próxima questão perguntava: 'tem alguém atrasado?'.
		2ºB	Desciam para a prova com *tablet* 16 alunos. Ficaram 15 alunos para a prova de papel, 32 presentes, um ausente. Na turma tem um aluno com deficiência com laudo e um aluno sem laudo, que se recusou a realizar a prova. Saiu da sala e a professora precisou solicitar um responsável para busca-lo. O aluno com laudo desceu para realizar a prova no *tablet*. O aluno K. retornou à sala e realizou a prova, aluno com problemas familiares, de aprendizagem e comportamento.
		2ºC	Professora não deu as instruções e já começou a prova. Primeiras questões (1, 2, 3, 4, 5): 'ah, que fácil'. Fala de aluno: 'O que é letra inicial?' (Item 6). Fala de aluno: 'O que é destacado?' (item 11). A partir do item 14 as crianças começaram a reclamar dos textos longos 'mais textos para ler?'. 'Nossa, quanto texto!'. 'Ah, eu já estou cansada da prova'. O professor não tinha o guia de aplicação.

ANEXO E – Controle do tempo (em minutos) de aplicação da versao papel e lápis e TBC da PB – Leitura, por Emef e por turma, 2016

Dia	Emef	Período	Total	Início	Término	Duração
25/out	**Moçambique**	MANHÃ	A-31	9:30	10:30	1:00
		MANHÃ	B-30	9:35	10:11	0:36
		MANHÃ	C-32	9:35	10:20	0:45
		Total	**93**			
27/out	**Costa do Marfim**	TARDE	A-28	14:33	15:11	0:38
		TARDE	B-28	14:30	15:15	0:45
		TARDE	C-27	14:30	14:57	0:27
		Total	**83**			
27/out	**República do Congo**	TARDE	A-32	16:50	17:14	0:24
		TARDE	B-30	16:55	17:44	0:49
		TARDE	C-32			
		TARDE	D-32	17:00	17:50	0:50
		Total	**126**			
28/out	**Etiópia**	MANHÃ	A-30	9:45	10:05	0:20
		MANHÃ	B-30	9:20	10:10	0:50
		MANHÃ	C-30	9:42	10:20	0:38
		Total	**90**			
28/out	**Angola**	TARDE	A-29	14:32	15:14	0:42
		TARDE	B-29	14:30	15:25	0:55
		TARDE	C-29	14:25	14:55	0:30
		Total	**87**			
31/out	**Libéria**	MANHÃ	A-30	9:35	10:05	0:30
		MANHÃ	B-30	9:32	10:06	0:34
		MANHÃ	C-30	9:27	10:17	0:50
		Total	**90**			
31/out	**Cabo Verde**	TARDE	A-30	14:15	14:55	0:40
		TARDE	B-31	14:09	14:58	0:49
		TARDE	C-30	14:09	14:49	0:40
		Total	**91**			
01/nov	**Argélia**	TARDE	A-30	14:20	14:45	0:25
		TARDE	B-28	14:31	15:06	0:35
		Total	**58**			
01/nov	**Líbia**	TARDE	A-23	16:30	17:15	0:45
		TARDE	B-23	16:37	17:12	0:35
		Total	**46**			
03/nov	**Benim**	TARDE	A-30			
		TARDE	B-30	13:57	14:23	0:26
		Total	**60**			

Continua

continuação

Dia	Emef	Período	Total	Início	Término	Duração
03/nov	**Egito**	TARDE	A-32	16:24	16:52	0:28
		TARDE	B-33	16:34	17:20	0:46
		TARDE	C-32	16:45	17:20	0:35
		Total	**97**			
04/nov	**Camarões**	TARDE	A-34	14:07	14:39	0:32
		TARDE	B-33	14:00	14:30	0:30
		TARDE	C-33			
		Total	**100**			
04/nov	**África do Sul**	TARDE	A-36	16:43	17:35	0:52
		TARDE	B-35			
		Total	**71**			
07/nov	**Ruanda**	MANHÃ	C-31	10:37	11:03	0:26
		TARDE	A-34	13:55	14:35	0:40
		TARDE	B-31	14:02	14:34	0:32
		Total	**96**			
08/nov	**Marrocos**	MANHÃ	A-32	10:31	11:15	0:44
		MANHÃ	B-32	10:24	11:15	0:51
		MANHÃ	C-32	10:30	11:09	0:39
		Total	**96**			
	MÉDIA					0:38
	DESVIO PADRÃO					0:09