Avaliação dos Subtestes de Leitura e Escrita do Teste de Desempenho Escolar através da Teoria de Resposta ao Item

Assessment of the Reading and Writing Subtests of the School Achievement Test through Item Response Theory

Luiza Feijó Knijnik[*, a], Cláudia Hofheinz Giacomoni[a], Cristian Zanon[b] & Lilian Milnitsky Stein[c]

[a]*Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brasil,* [b]*Universidade Federal de São Francisco, Itatiba, São Paulo, Brasil* & [c]*Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brasil*

Resumo

O Teste de Desempenho Escolar (TDE) avalia a aprendizagem escolar através de três subtestes: leitura, escrita e aritmética. O objetivo deste estudo foi conhecer quais partes do continuum de habilidades são medidas pelos subtestes de leitura e escrita e a quantidade de informação fornecida, utilizando a Teoria de Resposta ao Item (TRI). A amostra foi composta de 1850 crianças. Os resultados indicaram que o subteste escrita mensura precisamente níveis médios de habilidade e menos satisfatoriamente níveis baixos e altos. O subteste leitura revelou ter discriminação apropriada para níveis baixos e médios de habilidade. As partes do continuum de habilidade que estão sendo medidas adequadamente estão fornecendo alta quantidade de informação, demonstrando que os subtestes leitura e escrita estão funcionando bem.

Palavras-chave: Rendimento escolar, Teste de Desempenho Escolar, avaliação psicológica, teoria de resposta ao item.

Abstract

The School Achievement Test (SAT) assesses learning through three subtests: reading, writing and arithmetic. The goal of this study was to know which parts of the ability continuum are measured by the subtests as well as the amount of information provided in the reading and writing subtests using the Item Response Theory (IRT). The sample consisted of 1850 children. Results indicated that the writing subtest precisely measures average levels of ability and less satisfactorily low and high levels. The reading subtest revealed to have adequate discrimination for low and average levels of ability. The parts of the ability continuum which are adequately measured provide high amounts of information, indicating that reading and writing subtests are doing well.

Keywords: Educational achievement, School Achievement Test, psychological evaluation, item response theory.

A avaliação educacional é um tema de interesse para que metodologias eficazes de ensino sejam desenvolvidas, assim contribuindo para que cada vez mais pessoas possam ter acesso a uma educação de qualidade. Atualmente a preocupação com a avaliação educacional está refletida, por exemplo, via número crescente de ações do Ministério da Educação (MEC) como a criação do Sistema de Avaliação do Ensino Básico (SAEB), do Exame Nacional do Ensino Médio (ENEM) e do Exame Nacional de Desempenho de Estudantes (Enade; Sisto, Sbardelini, & Primi, 2000). O desenvolvimento e aprimoramento de medidas que possibilitem a verificação do desempenho dos alunos, como testes psicológicos e educacionais, portanto, surgem como ferramentas que podem auxiliar na qualificação tanto de professores, quanto de alunos.

A avaliação de desempenho escolar pode ser vista como uma forma de acompanhar a aprendizagem do aluno, um processo que permite a revisão de métodos de ensino e da aprendizagem, um estímulo ao trabalho didático e um constante apoio ao estudante. Possibilita, ainda, a identificação de pontos críticos que precisam ser redefinidos. A avaliação educacional através de testes de desempenho, por exemplo, podem servir como balizadores dos efeitos de programas de treinamento ou instrução específicos, obtendo os possíveis efeitos da aprendizagem (Anastasi

[*] Endereço para correspondência: Pós-Graduação em Psicologia – PUCRS, Av. Ipiranga, 6681, prédio 11, sala 940 Porto Alegre, RS, Brasil. CEP 90619-900. Fone: (51) 3353-7737. E-mail: luiza.knijnik@gmail.com, giacomon@uol.com.br, cristianzanon@yahoo.com.br e lilian@pucrs.br

& Urbina, 2000). Um levantamento da literatura (Knijnik, Giacomoni, & Stein, 2013) indica que muitos estudos brasileiros que analisaram o desempenho escolar de alunos do Ensino Fundamental, assim como suas relações com diversas outras variáveis, têm utilizado o Teste de Desempenho Escolar (TDE; Stein, 1994) como instrumento avaliador.

Teste de Desempenho Escolar

O TDE é o único instrumento psicopedagógico para avaliação ampla da aprendizagem, validado e normatizado para a população brasileira. Concebido para aplicação individual, o teste avalia de forma ampla as capacidades básicas para o desempenho escolar em três áreas específicas: leitura, escrita e aritmética. Ele foi normatizado com o objetivo de avaliar escolares de 1ª a 6ª séries do Ensino Fundamental.

O instrumento é composto por três subtestes organizados da seguinte forma: escrita do nome próprio e 34 itens que avaliam a habilidade de escrita – escrita de palavras contextualizadas, apresentadas sob a forma de ditado; 70 itens para avaliação da habilidade leitora – reconhecimento de palavras isoladas, e 35 itens que avaliam as habilidades em aritmética – solução oral de problemas e cálculo de operações aritméticas por escrito (Stein, 1994). Para a construção de cada um dos subtestes do TDE processos distintos foram desenvolvidos. Estudos de validade de conteúdo foram conduzidos, bem como análises de consistência interna para cada subteste, variando de 0,79 a 0,94. O coeficiente Alfa total foi 0,95 (para mais detalhes ver Knijnik et al., 2013).

Estudos Utilizando o TDE

Desde sua publicação o TDE não sofreu nenhuma atualização de conteúdo ou normas, tampouco foi lançado qualquer outro instrumento com os mesmos objetivos. Desta forma, o TDE tem sido largamente utilizado em estudos científicos, bem como ferramenta clínica para avaliação ampla do desempenho escolar por profissionais de diferentes áreas, como psicólogos, pedagogos, psicopedagogos e professores (Boscariol et al., 2011; A. A. Ferreira, Conte, & Marturano, 2011; Riechi, Moura-Ribeiro, & Ciasca, 2011; Silva & Beltrame, 2011).

No levantamento realizado por Knijnik et al. (2013) foram analisadas 222 publicações que utilizaram o TDE como medida de desempenho escolar, e poucos estudos apresentaram críticas quanto ao teste, ou propuseram alguma revisão. A maior parte dos estudos utilizou o teste como um instrumento satisfatório para seus objetivos, o que pode ser um indicativo de que o teste continua cumprindo seu propósito, e mantém sua qualidade, mesmo após 18 anos.

Ainda assim, mesmo que este levantamento indique que o TDE vem sendo amplamente utilizado, os padrões para testes psicológicos e educacionais da *American Psychological Association* indicam que um teste deve ser aperfeiçoado quando surgem novos dados de pesquisa ou ocorrem mudanças significativas na área que tornem o teste inadequado para uso. Sendo assim, é desejável que estudos psicométricos e atualizações de instrumentos sejam realizados periodicamente (Adams, 2000; *International Test Commission*, 2000/2003).

Neste sentido, algumas questões amplas podem ser colocadas a respeito do TDE, tais como, o teste segue medindo o que se propõe a medir? Os itens que compõe o teste são adequados para medir o que pretendem? Quão bem os itens dos subtestes discriminam o desempenho de diferentes alunos? Alguns incipientes estudos sobre o TDE vêm buscando responder de forma parcial aspectos dos questionamentos levantados, apontando lacunas do instrumento. Lúcio, Pinheiro e Nascimento (2009), por exemplo, investigaram o impacto da introdução de uma nova classe de erros no subteste de leitura do TDE na distribuição dos escores, considerando incorretas respostas de silabação (presença de pausas entre as sílabas ou letras) e correção espontânea, o que originalmente não é considerado erro na correção do teste. As autoras avaliaram 306 crianças de 1ª a 4ª séries pelo critério original do manual, o escore bruto 1 (EB1), assim como pelo novo critério proposto no estudo, escore bruto 2 (EB2). Foi observado que a adoção do novo critério EB2 tornou o teste mais discriminativo, indicando que o critério atual de correção é muito permissivo, ainda que a adoção do novo critério não tenha impedido o efeito de teto. As autoras apontam que este efeito seria consequência do excesso de itens fáceis e escassez de palavras difíceis no subteste, os quais seriam capazes de discriminar as habilidades das crianças mais experientes e dos leitores mais capazes. Salienta-se, entretanto, que a adoção do novo critério de correção EB2 não impediu o efeito de teto encontrado no estudo, indicando que mesmo este novo critério ainda não seria tão discriminativo quanto as autoras desejariam.

As mesmas autoras (Lúcio & Pinheiro, 2012), em outro estudo, realizaram ainda uma análise clássica dos itens do subteste de leitura do TDE, com 341 crianças de 2ª a 5ª série, utilizando os mesmos critérios de correção descritos no estudo anterior (o do manual – EB1, e o novo, EB2, mais rígido). Foi realizada uma seleção de itens do subteste a partir dos índices de discriminação (D, ou diferença entre os escores dos sujeitos que compõem os 27% mais habilidosos e os 27% menos habilidosos) e dificuldade (*p* ou proporção de acerto no teste), mantendo-se os itens com índice de discriminação (D) maior que 0,4. Com este ponto de corte foram selecionados apenas 20 itens do subteste utilizando o critério de acerto do manual EB1 (dificuldade entre 0,52 e 0,84; discriminação entre 0,41 e 0,84), não sendo analisado por isso o índice de dificuldade. Partindo do critério de acerto novo EB2 e considerando os índices de dificuldade (0,25 e 0,82) e discriminação (0,44 e 0,81) foram selecionados 25 itens. Desta forma 20 itens do subteste leitura foram selecionados a partir de análises estatísticas considerando EB1, e 25 itens foram selecionados a partir de EB2, dentre 70 itens que compõe o subteste. Através da análise de discriminação e dificuldade dos itens,

o estudo concluiu que o subteste de leitura demonstra problemas em discriminar as habilidades de leitores mais competentes, indicando que a maioria dos itens presentes no subteste de leitura não seria capaz de diferenciar com precisão as habilidades que pretende medir. As conclusões deste estudo corroboraram o estudo anterior (Lúcio et al., 2009) no sentido de apontar que a qualidade e distribuição dos itens de leitura deveriam ser revistos, pois o subteste possui muitos itens fáceis e poucos itens difíceis (Lúcio & Pinheiro, 2012).

O estudo de F. L. Ferreira et al. (2012) realizou a normatização dos subtestes de escrita e aritmética para uma amostra mineira de 1034 participantes. Os autores compararam as normas originais (de Porto Alegre, RS; Stein, 1994) com o desempenho de escolares de dois municípios de Minas Gerais, constatando diferenças significativas entre as amostras nos dois subtestes (magnitude de efeito *d* elevado para 2ª, 3ª, 5ª e 6ª série – acima de 0,8 – na escrita; e elevado – *d* = 0,88 – na 2ª série para aritmética). A amostra gaúcha apresentou um desempenho 20,05% superior em relação à amostra mineira, levando os autores a concluírem que as normas gaúchas não são adequadas para Minas Gerais. Neste sentido, o estudo sugeriu que as diferenças regionais quanto aos conteúdos curriculares presentes no Brasil demandam que testes de avaliação do desempenho escolar, como o TDE, devem ter normas regionalizadas. Este dado parece ir de encontro com a constatação feita no levantamento de Knijnik et al. (2013) de que o TDE vem sendo satisfatoriamente utilizado em estudos realizados em todas as regiões brasileiras, sem críticas quanto à regionalização das normas. Salienta-se que o estudo mineiro utilizou uma amostra proporcional não probabilística, que não permite a avaliação da representatividade da população, não sendo generalizável, enquanto a normatização original valeu-se de uma amostra probabilística aleatória e estratificada, que proporciona condições para avaliação da representatividade da amostra em uma população, fornecendo resultados potencialmente generalizáveis (Fowler, 1993).

Outro achado relevante de F. L. Ferreira et al. (2012) foi o desempenho 17,01% superior dos alunos da rede privada de ensino, em ambos os testes de escrita e aritmética, em relação aos alunos da rede pública. O estudo original (Stein, 1994), entretanto, não encontrou diferença entre o tipo de escola, assim como grande parte dos estudos que vêm utilizando o TDE (Knijnik et al., 2013). A diferença significativa destes resultados, no entanto, pode ser ocasionada pelo tamanho da amostra (1034) utilizada no estudo.

Os estudos apresentados apontam críticas em relação ao estado atual do teste, tanto em relação aos dados normativos, quanto aos critérios de correção. No entanto, nenhum destes estudos forneceu informações a respeito da qualidade dos subtestes do TDE. Tampouco em relação à discriminação das habilidades relativas ao aprendizado escolar dos participantes ao longo do continuum de habilidades que o teste se propõe a medir. Estas informações acerca das propriedades psicométricas do teste são importantes, pois podem auxiliar a compreender o quanto é necessário alterar o mesmo, e podem ser obtidas e analisadas através de dois principais vieses teóricos, descritos a seguir.

Teoria Clássica dos Testes e a Teoria de Resposta ao Item

De forma geral, existem duas abordagens básicas para avaliação de testes, uma clássica - a Teoria Clássica dos Testes (TCT), e outra que vem sendo estudada mais atualmente – a Teoria de Resposta ao Item (TRI). Um ponto a salientar em relação à TCT é que tanto a dificuldade quanto a discriminação de um item dependem da amostra particular a partir da qual os itens foram obtidos (Hambleton & Slater, 1997). Isto significa que seus valores podem se alterar quando calculados para amostras diferentes daquelas utilizadas na análise inicial de itens. Desta forma, as características do teste irão variar dependendo das pessoas que estiverem respondendo o mesmo (Urbina, 2007). Esta dependência da amostra resulta em um item se tornar mais fácil ou difícil (característica psicométrica de dificuldade) dependendo da aptidão do grupo de respondentes que se submeteu ao teste. Ou seja, o parâmetro de dificuldade do item pode variar em cada pesquisa, em função da amostra (Hambleton & Jodoin, 2003; Hambleton, Robin, & Xing, 2000; Lord & Novick, 1968). Isto é, na TCT os parâmetros dos itens variam em cada amostra, e os escores dos participantes dependem dos itens e da amostra (por exemplo, se um participante está dentro de uma amostra que tem ótimo desempenho, o seu desempenho será considerado ruim se for inferior ao da maioria). Por outro lado, ao se utilizar a TRI, a dificuldade dos itens não depende dos escores dos participantes, assim como os escores dos participantes não dependem nem dos itens, nem da amostra a qual fazem parte.

A TRI baseia-se na teoria do traço latente (habilidade ou aptidão) de Lazersfeld (Pasquali & Primi, 2003), que procura relacionar variáveis observáveis (desempenho em itens de um teste, por exemplo) e traços hipotéticos não observáveis ou aptidões que são responsáveis pelo surgimento das variáveis observáveis. A resposta que o indivíduo dá ao item depende do nível de habilidade que ele possui (Embretson, 1996). Os níveis de habilidade de uma pessoa podem ser avaliados dentro de um *continuum* que vai de muito pouca a muita habilidade. Geralmente, a escala usada apresenta média 0 e desvio-padrão 1, sendo que os extremos são -3 e +3. Assim, itens com dificuldade em entre -3 e -2 são considerados itens fáceis; itens com dificuldade em torno de 0 são médios e itens com dificuldade entre +2 e +3 são considerados difíceis. Outra vantagem da TRI é que o nível de habilidade dos sujeitos também é dado na mesma escala. Assim, sujeitos com habilidade em torno de 0 apresentam habilidade média, sujeitos com habilidade em torno de -2 apresentam habilidade baixa e sujeitos com habilidade acima de +2 apresentam habilidade alta.

A TRI é uma teoria estatística baseada em modelos matemáticos e que se propõe a prever o escore de uma

pessoa em um teste baseado no seu próprio traço latente ou habilidades; e estabelecer a relação entre o desempenho de uma pessoa em determinado item e o conjunto de traços subjacentes ao desempenho neste item através da curva característica do item (CCI). Na CCI, os examinandos com mais habilidade terão maior probabilidade de fornecer uma resposta correta do que aqueles com menos habilidade (Hambleton & Slater, 1997). Isto é possível devido à invariância de parâmetro dos itens e das habilidades, o que permite que os mesmos itens em amostras diferentes mantenham suas propriedades estatísticas, como sua dificuldade e discriminação, assim como que o escore de uma pessoa representando o seu traço latente não seja dependente dos itens de um teste (Hambleton, Swaminathan, & Rogers, 1991).

A aplicação da TRI parte de um pressuposto básico: a unidimensionalidade. O princípio da unidimensionalidade assume que há somente um construto sendo medido por vários itens, ou pelo menos uma aptidão (fator ou traço) dominante responsável pelo desempenho num conjunto de itens de um teste (Pasquali & Primi, 2003). Contudo, a busca pela unidimensionalidade pura pode ser inatingível, dado a complexidade dos construtos medidos. Por isso, o que está em jogo, geralmente, não é se há ou não unidimenisonalidade, mas o quanto há de unidimensionalidade em um conjunto de dados. Algumas formas avaliar evidências de unidimensionalidade são através de análise fatorial exploratória (AFE), análise fatorial confirmatória e análise paralela. Quando AFE são usadas, uma possibilidade é a verificação da diferença do primeiro autovalor extraído em relação ao segundo. Se o primeiro autovalor for pelo menos quatro vezes maior que o segundo, há evidências de unidimensionalidade (Hambleton et al., 1991).

De maneira ampla, as vantagens do uso da TRI têm a ver principalmente com a invariância dos parâmetros dos itens e com a produção de informação sobre o teste e seus itens. Ou seja, é possível obter detalhes a respeito de em que parte do traço latente um teste e seus itens estão fornecendo mais informações (Zanon, Hutz, Yoo, & Hambleton, 2012). A informação produzida por cada item é diretamente proporcional ao seu poder de discriminação, e a soma da informação produzida por todos os itens produz a informação total do teste. A inspeção da quantidade de informação produzida por um teste em partes específicas do continuum de habilidade permite que o pesquisador conheça com precisão onde o teste está medindo mais precisamente (ou seja, onde há menos erro de medida) o construto e onde a medição está mais deficitária. A relevância dessa inspeção deve-se a possibilidade de visualização clara e rápida de onde o teste mede melhor, onde mede pior e que itens podem ser acrescentados ou retirados para adequar o teste ao propósito do pesquisador. Por exemplo, se a medição está mais deficitária nos extremos (para sujeitos com baixos e altos níveis de habilidade), o pesquisador pode acrescentar itens fáceis e difíceis para suprir a imprecisão do teste. Essa inclusão de itens diminuirá o erro de medida nessas regiões do contínuo, o que aumentará a fidedignidade da medida. A avaliação da quantidade de informação produzida também pode ajudar na constatação de que a retirada de itens pode ser feita sem prejudicar a qualidade da medida. Isso pode acontecer quando se verifica muitos itens de dificuldade semelhante e com considerável poder de discriminação (Hambleton, 2005). Tendo em vista que no estudo original do TDE foram feitas análises clássicas, este tipo de análise ainda não foi realizada para subtestes do TDE, sendo possível somente com o uso da TRI.

Desta forma, o presente estudo focará na análise através da TRI de dois dos subtestes do TDE: leitura e escrita. Estes dois subtestes foram selecionados, sem a inclusão do subteste aritmética, devido à sua relação próxima e complementar. Alguns autores consideram que leitura e escrita fazem parte da mesma constelação de processos cognitivos, conectados, pois dependem da aquisição de conhecimentos similares, sendo seu desenvolvimento paralelo (Fitzgerald & Shanahan, 2000; Graham & Hebert, 2011). Estudos a respeito da ligação entre leitura e escrita têm focado suas análises nos componentes de cada uma destas habilidades para compreender os efeitos de uma sobre a outra (Cunha & Santos, 2006). A principal premissa é que melhores escritores seriam melhores leitores, e estudos têm demonstrado que há uma relação mais próxima entre o desempenho em leitura e escrita do que no desempenho em outras áreas acadêmicas (Fitzgerald & Shanahan, 2000). Desta forma, o presente estudo partilha desta visão de proximidade e reciprocidade das habilidades e leitura e escrita, e irá se dedicar somente a estes dois subtestes. Ressalta-se que o subteste de aritmética deverá também ser estudado em mais profundidade em um próximo estudo.

O objetivo principal deste estudo foi avaliar as propriedades psicométricas dos subtestes de leitura e escrita do TDE utilizando a TRI. Mais especificamente, objetivou-se: (a) avaliar a dimensionalidade dos subtestes, através de análises de eixos principais, (b) verificar que partes do continuum de habilidades os subtestes estão avaliando e (c) conhecer a quantidade de informação fornecida pelos subtestes nestas áreas.

Método

Amostra

A amostra foi composta por dados oriundos de pesquisas realizadas em quatro estados brasileiros. Os dados sociodemográficos da amostra são descritos na Tabela 1.

Procedimento de Coleta de Dados

A partir do levantamento acerca das publicações utilizando o TDE desde sua publicação (Knijnik et al., 2013) foi feito contato com os pesquisadores responsáveis pelas pesquisas encontradas a fim de solicitar autorização para utilização dos dados do TDE. Seguindo a orientação do Comitê de Ética para procedimentos no compartilhamento de dados, aqueles pesquisadores que concordaram em ce-

Tabela 1
Dados Sociodemográficos da Amostra

Características	Frequência (%)
Sexo	
Masculino	1007 (54,5)
Feminino	843 (45,5)
Série	
1ª	180 (9,7)
2ª	672 (36,3)
3ª	407 (22)
4ª	360 (19,5)
5ª	157 (8,5)
6ª	74 (4)
Origem dos dados	
SC	779 (42,1)
SP	520 (28,1)
RS	321 (17,4)
PB	230 (12,4)
Tipo de Escola	
Pública não especificada	170 (9,2)
Pública estadual	324 (17,5)
Pública municipal	1224 (66,2)
Particular	113 (6,1)
Abrigo*	7 (0,4)
Missing	12 (0,6)

Nota. * Os participantes oriundos de abrigos foram excluídos da amostra.

der seus dados assinaram um Termo de Compromisso de Dados, também assinado pelas pesquisadoras responsáveis pelo presente estudo.

Procedimento de Análise dos Dados

Os procedimentos de análise dos dados contaram com duas etapas. Inicialmente, partiu-se para avaliação da dimensionalidade dos subtestes, através de análises fatoriais. Uma vez que a TRI unidimensional pressupõe a unidimensionalidade do construto medido, conduziram-se análises de eixos principais para cada subteste no programa SPSS 19. As análises serão descritas em relação a cada subteste. Por fim, realizou-se análises de TRI para conhecer o quanto de informação é produzida no continuum de habilidade de cada subteste. Os itens do teste foram tratados como variáveis dicotômicas de acerto e erro (0 para erro, 1 para acerto).

As análises de TRI foram rodadas no programa BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) e o modelo escolhido foi o de dois parâmetros. Esse modelo foi selecionado, e não o de três parâmetros, por ser improvável que os testandos acertem itens por acaso, uma vez que o formato do subteste não fornece alternativas. Além disso, o seu ajuste aos dados foi adequado. Poucos itens apresentaram qui-quadrados significativos ($p < 0,01$). Como este teste é sensível ao tamanho da amostra procedeu-se à investigação dos resíduos padronizados dos itens individualmente e do teste como um todo no programa Resid-Plot (Liang, Han, & Hambleton, 2008). Esta forma de avaliar o ajuste do modelo, descrito em Hambleton e Jodoin (2003), requer a inspeção dos resíduos em diferentes pontos ao longo do contínuo de habilidades para cada item e para o teste. Resíduos padronizados próximos de zero indicam excelente ajuste e valores entre -2 e 2 indicam ajuste aceitável. A grande maioria dos resíduos avaliados estiveram entre -1 e 1. Além de sugerir ajuste aceitável, estes resultados indicam evidências de monotonicidade, ou seja, o valor da probabilidade de acertar as questões, estimado pela melhor curva característica do item, aumenta à medida que aumenta o nível de habilidade. O método de estimação usado foi o *maximum likelihood.*

Resultados e Discussão

Análise de Dados do Subteste Escrita

Com o intuito de avaliar a dimensionalidade do subteste de escrita, o conjunto de 34 itens foi submetido a uma análise de eixos principais usando como critérios de extração *eigenvalues* (autovalor) maiores que um e inspeção do *scree plot* (gráfico de sedimentação). O teste KMO foi de 0,98 e o teste de esfericidade de Bartlett foi significativo, o que indica adequação dos dados para a realização da análise. Inicialmente, foram verificados quatro fatores com autovalor maiores que um. O primeiro explicou 37% da variância e apresentou um autovalor de 12,58; o segundo explicou 6% da variância e apresentou um autovalor de 2,17; o terceiro explicou aproximadamente 3% da variância e apresentou um autovalor de 1,06; e o quarto explicou aproximadamente 3% da variância e apresentou um autovalor de 1,04. Uma vez que o objetivo dessa análise é verificar se há um fator predominante no conjunto de dados, outra análise foi rodada restringindo a extração a um único fator. Nessa solução, os itens apresentaram cargas fatoriais que variaram de 0,33 a 0,74. Além disso, verificou-se que o autovalor do primeiro fator é maior do que o segundo mais de quatro vezes. Essas evidências sugerem que há um fator predominante no grupo de itens, o que é condição necessária para avaliação do subteste através da TRI (Hambleton et al., 1991).

O foco deste trabalho é avaliar a adequação da medida do subteste como um todo ao longo do traço latente. Sendo assim, a descrição dos resultados centra-se na quantidade de informação e erros de mensuração produzidos pelo subteste de escrita (Figura 1).

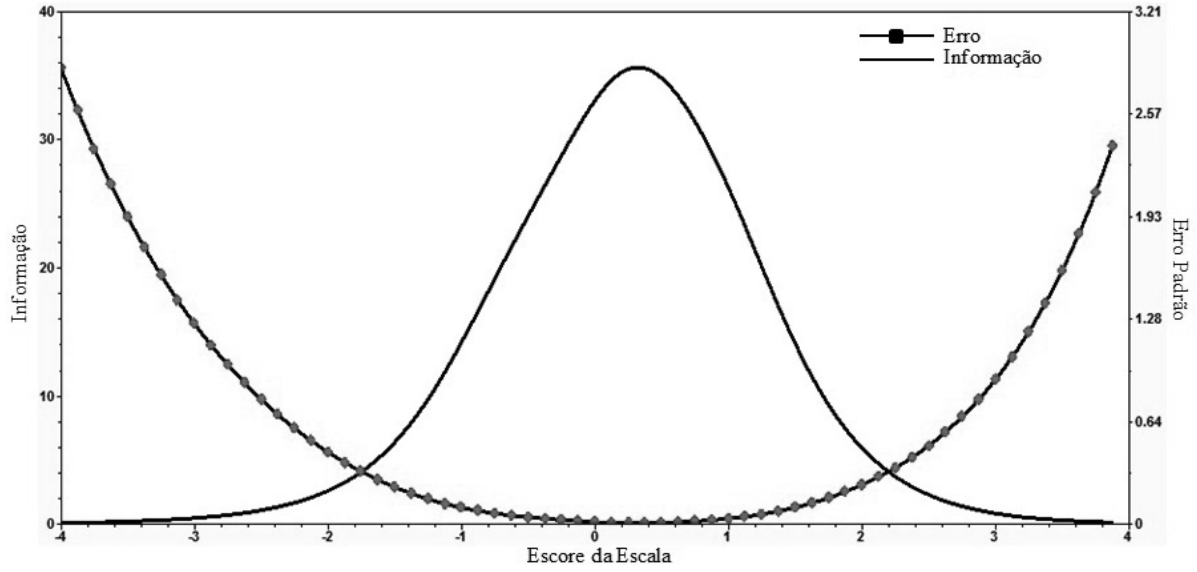

Figura 1. Curva de informação e erro padrão de medida produzida pelo conjunto de itens ao longo do nível de habilidade de escrita.

Tabela 2
Dificuldade e Discriminação dos Itens de Escrita

Item	Discriminação	Dificuldade	Item	Discriminação	Dificuldade
D01	1,520	-0,706	D18	1,270	0,287
D02	1,641	-0,691	D19	1,358	0,818
D03	0,717	-1,440	D20	1,109	0,269
D04	1,520	-0,848	D21	1,552	0,343
D05	1,814	-0,342	D22	1,138	0,361
D06	1,013	1,383	D23	1,732	0,564
D07	1,725	-0,495	D24	1,097	0,562
D08	2,061	0,046	D25	1,305	0,834
D09	2,062	0,227	D26	1,719	1,066
D10	1,230	-0,632	D27	1,091	0,842
D11	1,499	-0,532	D28	1,521	0,599
D12	1,166	-0,324	D29	1,898	1,004
D13	1,019	-1,326	D30	1,337	0,735
D14	1,632	-0,008	D31	1,788	0,486
D15	1,094	0,431	D32	1,309	1,609
D16	1,677	0,463	D33	0,964	0,843
D17	1,884	0,258	D34	1,675	0,997

A Figura 1 indica que o subteste de escrita está mensurando precisamente níveis médios de habilidade e menos satisfatoriamente níveis baixos e altos. A Tabela 2 apresenta a dificuldade (*M*=0,23; *DP*=0,74) e discriminação (*M*=1,4; *DP*=0,3) dos itens. Mais especificamente, o subteste discrimina adequadamente testandos com níveis de habilidade que variam aproximadamente entre -1,30 e 1,60 (ponto no qual se cruzam a quantidade de informação produzida e o erro padrão). Isso se deve a grande quantidade de informação produzida pelos itens nesse intervalo. Ademais, ressalta-se que 84% dos testandos dessa amostra encontram-se com escores dentro deste intervalo. Tendo em vista que a escala de habilidades tem seus extremos em -3 e +3, verificou-se que para valores maiores que 1,60 e menores que -1,30 o subteste de escrita apresenta maiores níveis de erro de mensuração, pois produz menor quantidade de informação, o que revela que o teste está medindo com menos precisão sujeitos que se encontram nessas regiões.

Análise de Dados do Subteste Leitura

Para avaliar a dimensionalidade do subteste de leitura, realizou-se uma análise de eixos principais com os seus 70 itens, utilizando como critérios de extração autovalores maiores que um e inspeção do gráfico de sedimentação. O teste KMO foi de 0,99 e o teste de esfericidade de Bartlett foi significativo, indicando adequação dos dados para a análise. Observou-se três fatores com autovalores maiores que um. O primeiro explicou 53% da variância e apresentou um autovalor de 39,33; o segundo explicou 6% da variância e apresentou um autovalor de 4,17; e o terceiro explicou aproximadamente 2% da variância e apresentou um autovalor de 1,72. Em seguida, foi rodada outra análise restringido a extração a um único fator. Nessa solução, os itens apresentaram cargas fatoriais que variaram de 0,48 a 0,88. Como se pode constatar, o autovalor do primeiro fator é mais de nove vezes maior que o do segundo, o que indica a predominância do primeiro fator no grupo de itens.

A Figura 2 apresenta a curva de informação e erro padrão de medida produzida pelo conjunto de itens ao longo do nível de habilidade de leitura. A Tabela 3 apresenta a dificuldade (*M*=-0,55; *DP*=0,53) e discriminação (*M*=2,36; *DP*=0,58) dos itens. Ressalta-se que os índices de discriminação destes itens são altos, o que contribui para avaliação e discriminação precisa dos participantes.

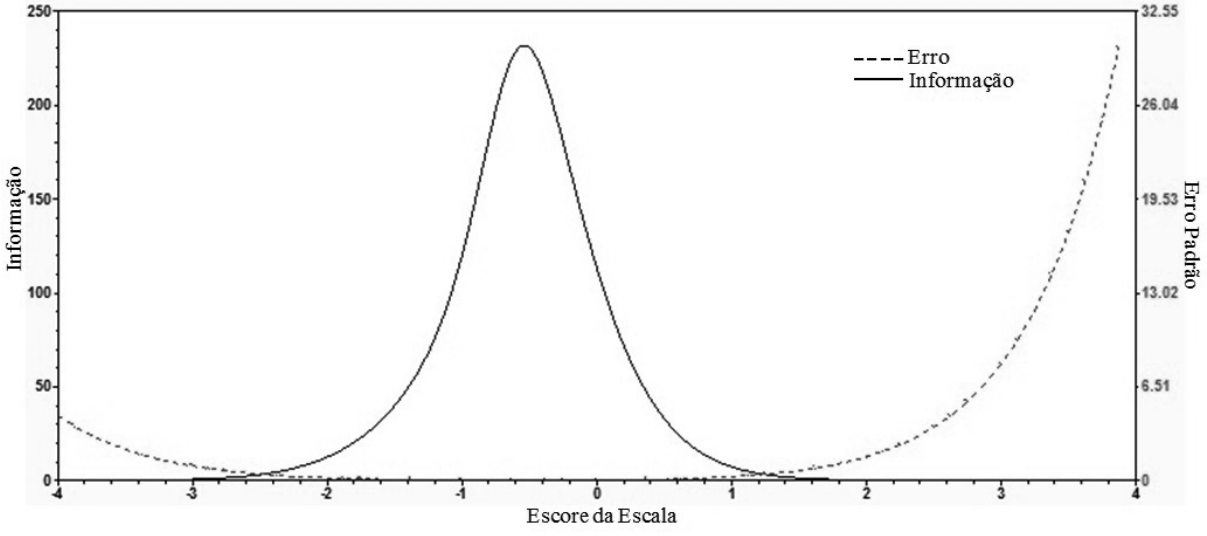

Figura 2. Curva de informação e erro padrão de medida produzida pelo conjunto de itens ao longo do nível de habilidade de leitura.

A análise de TRI para o subteste leitura revelou que há precisão apropriada para níveis baixos e médios de habilidade, sugerindo boa precisão de medida. Participantes com escores entre aproximadamente -2,15 e 0,85 são precisamente mensurados, o que corresponde a 77% dos testandos da amostra. Para valores acima e abaixo deste intervalo, o nível de mensuração do subteste é menor e há maiores níveis de erro de mensuração, o que revela que sujeitos com habilidade alta estão sendo avaliados mais precariamente.

Em suma, os resultados alcançados pelas análises realizadas indicam que ambos os subtestes têm poder de discriminação entre diferentes participantes. Algumas mudanças, entretanto, podem ser realizadas a fim de que o teste possa ficar ainda mais acurado em relação aos níveis de habilidade detec

Tabela 3
Dificuldade e Discriminação dos Itens de Leitura

Item	Discriminação	Dificuldade	Item	Discriminação	Dificuldade
01	1,842	-1,728	36	2,276	-0,302
02	1,645	-1,512	37	2,955	-0,608
03	1,637	-1,307	38	3,163	-0,519
04	2,010	-1,223	39	3,642	-0,542
05	1,734	-1,307	40	1,791	-0,453
06	1,372	-1,170	41	1,598	-0,003
07	1,691	-1,023	42	2,526	-0,378
08	2,524	-1,210	43	2,659	-0,459
09	2,580	-0,680	44	2,318	-0,317
10	1,622	-1,301	45	1,616	-0,216
11	1,677	-0,589	46	2,022	-0,716
12	1,813	-1,217	47	2,354	-0,418
13	2,022	-1,306	48	2,109	0,065
14	1,692	-1,073	49	2,399	-0,305
15	2,558	-0,596	50	3,664	-0,601
16	2,274	-0,680	51	2,978	-0,556
17	3,134	-0,766	52	2,935	-0,497
18	2,405	-1,099	53	1,892	-0,341
19	1,804	-0,510	54	1,626	-0,213
20	1,895	-0,369	55	2,724	-0,295
21	2,867	-0,612	56	2,419	-0,249
22	2,376	-0,429	57	2,861	-0,435
23	2,699	-0,740	58	2,102	0,227
24	2,967	-0,816	59	2,647	-0,494
25	2,763	-0,638	60	1,743	0,114
26	2,427	-0,900	61	2,763	-0,200
27	2,750	-0,594	62	2,193	-0,150
28	3,326	-0,696	63	2,853	-0,037
29	2,452	-0,398	64	2,277	-0,342
30	2,525	-0,526	65	1,758	-0,242
31	3,314	-0,451	66	1,896	0,086
32	3,283	-0,390	67	1,947	-0,167
33	3,748	-0,592	68	1,879	0,290
34	2,370	-0,161	69	1,777	-0,004
35	3,274	-0,723	70	2,025	0,074

Considerações Finais

A fim de avaliar as propriedades psicométricas dos subtestes leitura e escrita do TDE, o presente estudo utilizou um método de análise de dados estatisticamente robusto, a TRI. Alguns estudos, ainda que incipientes, buscaram analisar a qualidade do TDE e seus itens, entretanto, careceram de precisão quanto às informações que o tipo de análise realizada pode trazer, especificamente, quanto às áreas do continuum de habilidades que os subtestes estão avaliando.

Verificou-se que o subteste de escrita está medindo adequadamente os níveis médios de habilidade, mas não está mensurando bem níveis baixos e altos de habilidade. Isto pode estar prejudicando a diferenciação de alunos com pouca e alta habilidade dentro do continuum do traço latente. Desta forma, este subteste poderia ser aprimorado acrescentando-se itens mais difíceis e itens mais fáceis, para que possa fornecer mais informações nestas áreas.

Por outro lado, o subteste de leitura está discriminando satisfatoriamente níveis baixos e médios de habilidade, produzindo grande quantidade de informação nestas áreas. O nível de erro de mensuração é maior no traço de habilidade alto, sendo assim, itens mais difíceis devem ser acrescentados neste subteste para que crianças com maior habilidade possam ser mais bem discriminadas. Além disso, itens redundantes podem ser retirados deste subteste, pois a quantidade de informação produzida nos níveis baixo e médio de habilidade é grande. Este achado vai ao encontro das conclusões de Lúcio e Pinheiro (2012) e Lúcio et al. (2009), que apontaram a necessidade de inclusão de itens mais difíceis no subteste leitura do TDE.

Apesar de apresentar lacunas, as parte do continuum de habilidade que estão sendo medidas adequadamente estão fornecendo alta quantidade de informação, demonstrando que os subtestes de leitura e escrita estão funcionando bem nestes quesitos. Tal achado é corroborado pelos dados encontrados no levantamento de Knijnik et al. (2013), demonstrando que os subtestes do TDE são utilizados satisfatoriamente como medida de desempenho escolar em diversos estudos. Entretanto, levanta-se o questionamento quanto à precisão dos dados fornecidos pelas avaliações das habilidades de leitura e escrita realizadas através do TDE, visto que leitores muito habilidosos em leitura, por exemplo, não estão a par, sendo bem diferenciados de leitores com menos habilidade.

Desta forma, apesar da ampla utilização do TDE em pesquisas, o presente estudo apresentou evidências de que os subtestes de leitura e escrita devem ser aprimorados, e indicou possíveis caminhos que devem ser seguidos para este propósito. É necessário que o subteste de aritmética passe por similar avaliação, a fim de se obter informações a respeito do estado atual do instrumento completo. Sugere-se que futuros estudos realizem o mesmo tipo de análise aqui apresentada para o subteste de aritmética. Análises TRI por item de cada subteste também devem ser realizadas, uma vez que o propósito do presente estudo era caracterizar os subtestes de forma geral, não abarcando estas análises mais específicas.

De maneira ampla, os achados deste estudo são favoráveis à construção de uma versão revisada do TDE, com a atualização de seus subtestes e também de suas normas. Aspectos mais abrangentes como a adaptação do teste para a nova realidade do Ensino Fundamental de nove anos devem ser considerados nessa revisão. Outro aspecto que pode ser levado em conta na reformulação do TDE diz respeito à avaliação de diferenças entre o desempenho de alunos da rede pública e privada, conforme apontado por F. L. Ferreira et al. (2012). É possível hipotetizar que as diferenças entre estes dois níveis de ensino tenham aumentado desde a publicação das normas originais, em 1994 (quando não foi observada diferença), e que normas distintas sejam necessárias.

A TRI é um método que permite melhorar testes educacionais e psicológicos, pois permite ampla compreensão do que está sendo medido pelos instrumentos. Exemplo disso é o uso atual da TRI no Exame Nacional do Ensino Médio (ENEM), como forma de evitar que alunos que acertam questões através do "chute" tenham a mesma avaliação de alunos que de fato tem o conhecimento para acertar a questão. Desta forma, dois alunos podem acertar o mesmo número de questões, mas terão pontuações diferentes em função das questões específicas acertadas. Tendo em vista que a TRI pode qualificar certos instrumentos de avaliação, sugere-se que os procedimentos desenvolvidos neste estudo sejam adotados para avaliação de outros instrumentos psicométricos, a fim de aprimorar e qualificar os testes utilizados no cenário de avaliação educacional e psicológica brasileiro.

Referências

Adams, K. M. (2000). Practical and ethical issues pertaining to test revisions. *Psychological Assessment, 12*(3), 281-286.

Anastasi, A., & Urbina, S. (2000). *Testagem psicológica*. Porto Alegre, RS: Artmed.

Boscariol, M., Guimarães, C. A., Hage, S. R. V., Garcia, V. L., Schmutzler, K. M. R., Cendes, F., & Guerreiro, M. M. (2011). Auditory processing disorder in patients with language-learning impairment and correlation with malformation of cortical development. *Brain & Development*, *33*, 824-831.

Cunha, N. B., & Santos, A. A. A. (2006). Relação entre a compreensão da leitura e a produção escrita em universitários. *Psicologia: Reflexão e Crítica, 19*(2), 237-245.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341-349.

Ferreira, A. A., Conte, K. M., & Marturano, E. M. (2011). Meninos com queixa escolar: Autopercepções, desempenho e comportamento. *Estudos de Psicologia* (Campinas)*, 28*(4), 443-451.

Ferreira, F. L., Costa, D. S., Micheli, L. R., Oliveira, L. F., Pinheiro-Chagas, P., & Haase, V G. (2012). School Achievement Test: Normative data for a representative sample of elementary school children. *Psychology & Neuroscience, 5*(2), 157-164.

Fitzgerald, J., & Shanahan, T. (2000). Reading and writing relations and their development. *Educational Psychologist, 35*(1), 39-50.

Fowler, F. J. (1993). Survey research methods. *Applied Social Research Methods Series, 1*.

Graham, S., & Hebert, M. (2011). Writing to read: A meta-analysis of the impact of writing and writing instruction on reading. *Harvard Educational Review*, *81*(4), 710-785.

Hambleton, R. K. (2005). Applications of item response theory to improve health outcomes assessment: Developing item banks, linking instruments, and computer-adaptive testing. In J. Lipscomb, C. C. Gotay, & C. Snyder (Eds.), *Outcomes assessment in cancer* (pp. 445-464). Cambridge, UK: Cambridge University Press.

Hambleton, R. K., & Jodoin, M. (2003). Item response theory: Models and features. In R. F. Ballesteros (Ed.), *Encyclopedia of psychological assessment* (pp. 509-514). London: Sage.

Hambleton, R. K., Robin, F., & Xing, D. (2000). Item response models for the analysis of educational and psychological test data. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 553-585). San Diego, CA: Academic Press.

Hambleton, R. K., & Slater, S. C. (1997). Item response theory models and testing practices: Current international status and future directions. *European Journal of Psychological Assessment, 13*(1), 21-28.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Thousand Oaks, CA: Sage.

International Test Commission. (2003). *Diretrizes para o uso de testes: International Test Commission* (Instituto Brasileiro de Avaliação Psicológica, Trad.). (Original publicado em 2000)

Liang, T., Han, K. T., & Hambleton, R. K. (2008). *User's guide for ResidPlots-2: Computer software for IRT graphical residual analyses* (Version 2.0, Center for Educational Assessment Research Report no. 688) [Computer and manual software]. Amherst, MA: Center for Educational Assessment, University of Massachusetts.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores (with contributions by Allen Birnbaum)*. Reading, MA: Addison-Wesley.

Lúcio, P. S., & Pinheiro, A. M. V. (2012). *Novos estudos psicométricos para o subteste de leitura do Teste de Desempenho Escolar.* Manuscrito submetido para publicação.

Lúcio, P. S., Pinheiro, A. M. V., & Nascimento, E. (2009). O impacto da mudança no critério de acerto na distribuição dos escores do subteste de leitura do Teste de Desempenho Escolar. *Psicologia em Estudo* (Maringá)*, 14*(3), 593-601.

Knijnik, L. F., Giacomoni, C. H., & Stein, L. M. (2013). Teste de Desempenho Escolar: Um estudo de levantamento. *Psico-USF, 18,* 407-416.

Pasquali, L., & Primi, R. (2003). Fundamentos da Teoria da Resposta ao Item – TRI. *Avaliação Psicológica, 2*(2), 99-110.

Riechi, T. I. J., Moura-Ribeiro, M. V. L., & Ciasca, S. M. (2011). Impact of preterm birth and low birth weight on the cognition, behavior and learning of school-age children. *Revista Paulista de Pediatria*, *29*(4), 495-501.

Silva, J., & Beltrame, T. S. (2011, abr./jun.). Desempenho motor e dificuldades de aprendizagem em escolares com idades entre 7 e 10 anos. *Motricidade* (Santa Maria da Feira), *7*(2), 57-68.

Sisto, F. F., Sbardelini, E. T. B., & Primi, R. (Eds.). (2000). *Contextos e questões da avaliação psicológica.* São Paulo, SP: Casa do Psicólogo.

Stein, L. M. (1994). *TDE - Teste de desempenho escolar: Manual para aplicação e interpretação.* São Paulo, SP: Casa do Psicólogo.

Urbina, S. (2007). *Fundamentos da testagem psicológica.* Porto Alegre, RS: Artmed.

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2012). *A comprehensible application of item response theory to psychological test development*. Manuscript in preparation.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and Test Maintenance for Binary Items* (Versão 3.0) [Computer software]. Chicago, IL: Scientific Software.

Recebido: 17/01/2013
1ª revisão: 17/04/2013
2ª revisão: 27/05/2013
Aceite final: 12/06/2013