# ARTIGOS

# A aplicabilidade da Teoria de Resposta ao Item nos processos relacionados à avaliação de desempenho de estudantes de educação básica

Josué Sérgio de Souza*
Lukelly Fernanda Amaral Gonçalves**
Ramon Silva Ferreira***
Luana Lopes dos Santos Alves****
Luiz Carlos de Paiva*****
Colaboradores******

Resumo: A Teoria de Resposta ao Item (TRI) tem sido amplamente utilizada nas avaliações de larga escala no Brasil. No presente artigo desejou-se apresentar conceitos, pressupostos e modelos da TRI. A partir de levantamento bibliográfico, objetivou-se mostrar as vantagens dessa teoria para elaborar testes educacionais mais válidos e determinar resultados mais confiáveis para as avaliações de larga escala. Desse modo, a análise da TRI no presente documento não busca esgotar o assunto, mas auxiliar gestores na tomada de decisões relacionadas à avaliação de larga escala, tendo em vista o caso particular do eixo avaliação de desempenho do Sistema Permanente de Avaliação Educacional da Secretaria de Estado de Educação do Distrito Federal (SIPAEDF).

Palavras-chave: Teoria de Resposta ao Item. Avaliação de larga escala. SIPAEDF.

* Mestre em Matemática pela Universidade de Brasília (UnB). Professor da Secretaria de Estado de Educação do Distrito Federal (SEEDF) e atua na GAAP/DIAV/SUPLAV/SEEDF. Contato: josuesergio@uol.com.br
** Mestre em Educação pela Universidade de Brasília (UnB). Professora da Secretaria de Estado de Educação do Distrito Federal (SEEDF) e atua na GAAP/DIAV/SUPLAV/SEEDF. Contato: lukellyf@hotmail.com
*** Especialista em Avaliação Educacional (Faculdade Unyleva). Professor da Secretaria de Estado de Educação do Distrito Federal (SEEDF) e atua na GAAP/DIAV/SUPLAV/SEEDF. Contato: ramon.fisica@gmail.com
**** Mestre em Matemática pela Universidade de Brasília (UnB). Professora da Secretaria de Estado de Educação do Distrito Federal (SEEDF) e atua na GAAP/DIAV/SUPLAV/SEEDF
Contato: luanamatematica@hotmail.com
***** Mestre em Educação Profissional e Tecnológica pelo Instituto Federal de Goiás (IFG). Professor da Secretaria de Estado de Educação do Distrito Federal (SEEDF) e atua na GAAP/DIAV/SUPLAV/SEEDF. Contato: luizcarlos@edu.se.df.gov.br
****** Michelle Cruz Camargo de Oliveira; Juliana Martins Asevedo; Karine Rocha Lemes Silva.

## Introdução

A meta 7.22 do Plano Distrital de Educação (PDE) (estabelecido pela Lei n° 5.499, de 14 de julho de 2015) versa sobre a criação e implementação do Sistema Permanente de Avaliação Educacional do Distrito Federal. De modo a atender a tal demanda, foi publicada a Portaria n° 38, de 18 de fevereiro de 2020 (DISTRITO FEDERAL, 2020), para normatizar o Sistema Permanente de Avaliação Educacional da Secretaria de Estado de Educação do Distrito Federal (SIPAEDF). O SIPAEDF tem por finalidade "contribuir com a garantia da qualidade de educação do Distrito Federal, (re) direcionar políticas públicas educacionais e promover subsídios para intervenções pedagógicas e administrativas" (DISTRITO FEDERAL, 2020). A gestão do SIPAEDF é de responsabilidade da Diretoria de Avaliação (DIAV), vinculada à Subsecretaria de Planejamento, Acompanhamento e Avaliação (SUPLAV) da Secretaria de Estado de Educação do Distrito Federal (SEEDF). O SIPAEDF é constituído de dois eixos, a saber: avaliação de desempenho dos estudantes e avaliação de contexto. A avaliação de desempenho dos estudantes se dá por meio da aplicação da Prova DF, instrumento constituído por itens de múltipla escolha de Língua Portuguesa e Matemática. Cabe, portanto, à DIAV definir o método de avaliação mais adequado de modo a representar de forma mais fiel possível os resultados da aprendizagem, que seja válido, fidedigno e que não privilegie grupos específicos de estudantes. Além disso, o método escolhido deve permitir a comparação do desempenho de estudantes avaliados em épocas diferentes, já que o SIPAEDF é estruturado em ciclos e o primeiro vai de 2020 a 2025 . Nesse cenário, é necessária a reflexão por parte dos gestores da SEEDF sobre a metodologia a ser utilizada para gerar os resultados da Prova DF.

A necessidade de escolha de metodologia para verificar o desempenho de estudantes de uma rede de ensino não é exclusividade da DIAV. Diversos estados e municípios do Brasil realizam avaliação de desempenho de seus estudantes tendo em vista a promoção de qualidade da educação e o direcionamento de políticas públicas (GONÇALVES, et al., 2020). Logo, o material exposto neste trabalho se aplica não apenas ao SIPAEDF, mas a todo sistema de avaliação de larga escala que se propõe a avaliar o desempenho de seus estudantes por meio de itens de múltipla escolha.

Os processos de avaliação e seleção de indivíduos se dão, tradicionalmente, por resultados obtidos em provas, apresentados apenas por seus escores brutos ou padronizados. Dentro do que se entende por Psicometria Clássica, ou Teoria Clássica do Teste (TCT), as análises e interpretações estão sempre associadas à prova como um todo, o que torna inviável a comparação entre respondentes que não foram submetidos às mesmas provas. Para sanar essa e outras dificuldades da TCT, especialistas em Psicometria desenvolveram um conjunto de modelos matemáticos que a complementam em algumas de suas fragilidades. Tal modelo é conhecido como Teoria de Resposta ao Item (TRI). Entre as suas principais virtudes está permitir a comparação entre indivíduos que realizam duas provas totalmente diferentes. Outro ponto importante é que dois indivíduos com a mesma quantidade de acertos não necessariamente terão o mesmo escore final, já que a proficiência leva em consideração a coerência das respostas, tornando o resultado mais fidedigno. Essas e outras virtudes fazem com que o uso da TRI, em conjunto com a TCT, esteja em expansão nos exames e nas avaliações de larga escala no Brasil. Seu uso mais evidente perante a população brasileira é no Exame Nacional do Ensino Médio (Enem), exame sob a responsabilidade do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), que se tornou a principal porta de acesso ao ensino superior aos concluintes do ensino médio. O Inep definiu o uso da TRI no Enem para o cálculo da proficiência dos participantes a partir da edição de 2009 e, no ano seguinte, as vantagens da metodologia se mostraram necessárias para o tratamento isonômico dos participantes sem elevar muito os custos da aplicação após problemas de logística.

A edição do Enem do ano de 2010 passou por uma batalha judicial motivada por problemas na sua aplicação . Um lote de provas do 1º dia de aplicação, 6 de novembro de 2010, apresentava erros de impressão, tais como: gabaritos invertidos, cadernos de provas incompletos ou itens repetidos. A juíza Karla de Almeida Miranda Maia, da 7ª Vara Federal do Ceará, acatou o argumento do Ministério Público Federal (MPF) e determinou a imediata suspensão do Enem 2010 em todo o Brasil. A Justiça entendeu que o erro de impressão das provas levou prejuízo aos candidatos. O MPF, a Ordem dos Advogados do Brasil e outras entidades discutiam, inclusive, a anulação do exame, que contava com a participação de 3,3 milhões de participantes . Para interromper essa suspensão, a Advocacia Geral da União (AGU) protocolou recurso contra a decisão e uma petição com informações técnicas, baseando-se em dois argumentos: capacidade do Estado em solucionar o problema e garantia de isonomia para os estudantes que participassem de uma nova aplicação. Tal garantia de isonomia entre participantes da primeira e da segunda aplicação advém da Teoria de Resposta ao Item (TRI), conjunto de modelos matemáticos adotado

em 2009 no referido exame e que permite elaborar provas diferentes com o mesmo grau de dificuldade. O Enem sofria por falta de credibilidade devido aos problemas na sua aplicação do ano anterior e isso refletia em questionamentos a respeito da garantia de isonomia. Foi necessária até mesmo a manifestação da ONU para atestar a confiabilidade da metodologia para reaplicar a prova apenas para os participantes prejudicados pelas falhas de impressão. Por fim, o Enem 2010 foi retomado, com a reaplicação para cerca de dois mil participantes e utilizado por diversos estudantes na época como acesso ao tão sonhado curso superior.

Pouco mais de dez anos depois, a TRI segue utilizada no Enem sem os questionamentos técnicos de outrora. Mais do que isso, o seu uso continua em expansão para a realização de exames e de avaliação de desempenho por todo o Brasil. A TRI tem sido amplamente utilizada nos diversos processos referentes ao desenvolvimento de testes, tais como: elaboração de testes de larga escala, calibração de itens e construção de escalas de habilidades e de bancos de itens. Provas como o Sistema Nacional de Avaliação da Educação Básica (Saeb) e o Sistema de Avaliação do Rendimento Escolar do Estado de São Paulo (Saresp) utilizam a TRI em seus processos.

O presente trabalho tem como objetivo expor os conceitos, pressupostos e aplicações da TRI em avaliações educacionais de larga escala. Tem o escopo de fundamentar a escolha pela TRI, em conjunto com a TCT, nos processos relacionados ao eixo avaliação de desempenho dos estudantes do SIPAEDF, da concepção dos cadernos de prova à construção da escala de proficiência.

Para tanto, compreende-se, conforme Ghedin e Franco (2011, p. 107), que

> a metodologia deve ser concebida como um processo que organiza cientificamente todo o movimento reflexivo, do sujeito ao empírico e deste ao concreto, até a organização de novos conhecimentos, que permitam nova leitura/compreensão/interpretação do empírico inicial.

Sendo assim, enxergamos como de extrema importância um levantamento teórico sobre a importância da adoção da TRI em avaliações em larga escala.

Desse modo, para alcançar o objetivo geral do estudo, adotou-se a abordagem qualitativa, que é considerada um estudo dinâmico, flexível o qual "não se preocupa com representatividade numérica, mas, sim, com o aprofundamento da compreensão de um grupo social, de uma organização, etc." (SILVEIRA; CORDOVA, 2009, p. 31).

Já a natureza do estudo é a pesquisa básica, a qual objetiva "gerar conhecimentos novos, úteis para o avanço da Ciência, sem aplicação prática prevista. Envolve verdades e interesses universais" (SILVEIRA; CORDOVA, 2009, p. 34), visto que a continuidade empírica pode se dar a partir deste estudo preliminar.

Em suma, este é um trabalho do tipo exploratório, o qual proporciona "maior familiaridade com o problema, com vistas a torná-lo mais explícito ou a construir hipóteses" (SILVEIRA; CORDOVA, 2009, p. 35).

Quanto aos procedimentos adotados, nos valemos da pesquisa bibliográfica, a qual ocorreu a partir de busca em periódicos e bancos de teses e dissertações.

## 1. Noções Iniciais: conceitos relacionados à TRI no contexto de avaliação de larga escala

Construir um instrumento para medir uma variável nas ciências sociais abrange uma série de etapas que devem ser seguidas rigorosamente, tais como a conceituação dos comportamentos que definem operacionalmente o construto em questão, a elaboração de itens que acessem o construto, a administração dos itens elaborados para amostras pré-definidas, o refinamento do instrumento baseado em análise dos itens e a realização de estudos de validade e confiabilidade (ANDRADE; LAROS; GOUVEIA, 2010). Essas etapas são necessárias para se garantir que os escores em um instrumento sejam consistentes e realmente acessem o construto que se pretende avaliar (MATHISON, 2005).

Psicometria é uma área da Psicologia que consiste em técnicas utilizadas para quantificar um conjunto de comportamentos que se deseja conhecer melhor. Segundo Pasquali (2003), a Psicometria caracteriza-se por expressar numericamente um fenômeno psicológico. Para ele, a Psicometria fundamenta-se na teoria da medida em ciências para explicar o sentido que têm as respostas dadas pelo sujeito a uma série de tarefas e propor técnicas de medida dos processos mentais. É um ramo especializado da Psicologia que se dedica ao estudo e elaboração de testes de avaliação psicológica e ao desenvolvimento e aplicação dos conhecimentos estatísticos e de outros processos matemáticos à Psicologia. É uma área da Psicologia com uma concepção estatística que explica os comportamentos e aptidões por meio de testes cuja mensuração é feita através das respostas que os indivíduos fornecem a uma série de tarefas, tipicamente chamadas de itens. Para atingir esse objetivo, a Psicometria se vale, principalmente, de duas teorias: a Teoria Clássica dos Testes (TCT) e a Teoria de Resposta ao Item (TRI).

O modelo clássico da Psicometria tradicional (PASQUALI, 2003) está fundamentado na Teoria Clássica dos Testes (TCT). Esta tem como objetivo a

determinação das propriedades ou dos parâmetros do teste como um todo. As aptidões são medidas, em geral, pela soma das pontuações referentes às respostas dadas a uma série de itens, expressa no chamado escore total, revelando que o objetivo é explicar o resultado final total. No entanto, os procedimentos baseados no modelo clássico da Psicometria apresentam limitações que se refletem na qualidade dos testes, e a TRI surgiu no auge do aprimoramento de tais procedimentos.

Pasquali (2003), citando Thurstone (1959), assevera que um instrumento de medida não pode ser afetado pelo objeto de medida, pois esta influência limita ou prejudica a validade do instrumento. A TCT, apesar de bem fundamentada já nos anos 1950, continha o problema do instrumento construído dependendo intrinsecamente do objeto medido. Em outros termos, o teste será considerado fácil, mediano ou difícil a depender da aptidão do grupo de indivíduos que se sujeitou ao teste. Além disso, indivíduos que acertam a mesma quantidade de itens, mas com propriedades psicométricas distintas, apresentam o mesmo escore total (ANDRADE; TAVARES; VALLE, 2000).

A independência do instrumento de medida em relação ao objeto medido nos testes de inteligências é um dos motivos para o surgimento da TRI. Embora seja uma teoria que surgiu por volta da década de 50 do século passado, foi nos anos 80 que começou a ser difundida, com o avanço do desenvolvimento de softwares para uso prático dos algoritmos (RABELO, 2013). Portanto, a TRI preocupa-se com o estudo das características métricas dos itens, utilizando, para tanto, uma escala microscópica. Já a TCT tem seu foco direcionado ao próprio instrumento de medida e emprega, para tal, uma escala macroscópica (ANDRIOLA, 2009). Enquanto a TCT tem interesse em produzir testes de qualidade, a TRI se interessa por produzir tarefas (itens) de qualidade.

A TRI é um conjunto de modelos matemáticos que procuram representar a probabilidade de um indivíduo dar a resposta a um item como função dos parâmetros do item, e da habilidade ou proficiência dos respondentes. Os modelos relacionam variáveis observáveis (respostas aos itens de um teste) com aptidões não observáveis e que são responsáveis pelas respostas dadas pelo indivíduo (as quais se denominam traços latentes ou construtos). De acordo com essa relação, quanto maior a habilidade, maior a probabilidade de acerto do item.

O objetivo é estimar o nível de aptidão, o construto, do indivíduo a partir da análise das respostas dadas por ele a um conjunto de questões ou itens e para isso recorre-se a modelos matemáticos que relacionam as variáveis envolvidas nessa situação. No contexto da avaliação educacional, os traços latentes são as competências cognitivas dos estudantes e a TRI sugere formas de representar a relação entre a probabilidade de um estudante responder corretamente a um item e seus traços latentes ou habilidade na área de conhecimento avaliada.

A TRI é a abordagem que se concentra na relação entre a resposta de um indivíduo a um item de um teste e a posição desse indivíduo, em termos probabilísticos, da aptidão que está sendo medida, ou seja, possui como foco o estudo individualizado dos itens de um teste. Na TRI são medidos os traços latentes, que são características intrínsecas dos indivíduos que não podem ser medidas diretamente, isto é, não podem ser observadas fisicamente (COHEN; SWERDLIK; STURMAN, 2014).

Pasquali (2003) indica dois postulados básicos relacionados à TRI. O primeiro é que o desempenho do sujeito em uma tarefa pode ser predito a partir de um conjunto de fatores ou variáveis hipotéticas, a quem se denomina aptidões, traço latente ou construto e geralmente identificado pela letra grega theta (θ). O segundo trata da descrição da relação entre o desempenho e o traço latente por meio de uma equação matemática, chamada curva característica do item (CCI).

A TRI é amplamente utilizada em avaliações educacionais, em especial nas avaliações de larga escala. Ela permite estabelecer escalas de proficiência interpretáveis que possibilitam a comparação de indivíduos e o acompanhamento da evolução dos sistemas de ensino ao longo dos anos.

Dentro do arcabouço teórico da TRI há diversos modelos propostos que dependem essencialmente de três fatores (ANDRADE; TAVARES; VALLE, 2000). O primeiro é relacionado à natureza do item, se é dicotômico (apenas duas opções de resposta, certo ou errado) ou não dicotômico (item aberto). O segundo é relacionado com o número de populações envolvidas, se apenas uma ou mais de uma. Por fim, a quantidade de traços latentes que está sendo medida, se apenas um (modelo unidimensional) ou mais de um (modelo multidimensional). Neste trabalho, foca-se nos modelos para itens dicotômicos e unidimensionais, que coincide com o modelo mais utilizado no país nas avaliações de larga escala.

Há três tipos de modelos logísticos para itens dicotômicos, que se diferenciam pela quantidade de itens que utilizam para descrever o item. O modelo logístico de um parâmetro considera apenas a dificuldade. O modelo logístico de dois parâmetros considera a dificuldade e a discriminação. Por fim, o modelo logístico de três parâmetros considera a

dificuldade, a discriminação e a probabilidade de acerto ao acaso, que é o modelo mais comum e utilizado no Enem e no Saeb.

Matematicamente, o modelo logístico de três parâmetros pode ser descrito como:

$$P\left(U_{ij}=1|\theta_j\right)=c_i+(1-c_i)\frac{1}{1+e^{-Da_i\left(\theta_j-b_i\right)}}$$

com i = 1, 2, ..., I e j = 1, 2, …, n, onde

- Uij é uma variável dicotômica que assume os valores 1, quando o indivíduo j responde corretamente o item i ou zero quando o indivíduo j não responde corretamente ao item i;
- 0j representa a habilidade (traço latente) do j-ésimo indivíduo;
- P(Uij = 1| 0j ) é a probabilidade de um indivíduo j com a habilidade 0j responder corretamente o item i e é chamada de Função de Resposta ao Item - FRI;
- bi é o parâmetro de dificuldade do item i, medido na mesma escala da habilidade;
- ai é o parâmetro de discriminação do item i, com valor proporcional à inclinação da Curva Característica do Item (CCI) no ponto bi;
- ci é o parâmetro do item que representa a probabilidade de indivíduos com baixa habilidade responderem corretamente ao item i (probabilidade de acerto casual);
- D corresponde a um fator de escala, constante e igual a um. Quando se desejar que a função logística forneça resultados semelhantes aos da função ogiva normal, utiliza-se para o fator D o valor 1,7.

A relação expressa pela equação acima está descrita na Curva Característica do Item (CCI). A CCI informa as diferentes probabilidades de acerto que diferentes sujeitos com valores diferentes de variáveis latentes 0 apresentam. O valor de 0 varia de menos infinito a mais infinito e, por sua vez, o valor de P (0) varia de zero a um. O gráfico é uma sigmóide, curva em forma de "s", com duas assíntotas horizontais, cujo formato fornece diversas informações importantes sobre a qualidade do item a que se refere.

O modelo baseia-se no fato de que indivíduos com maior probabilidade de acertar o item possuem maior proficiência e que essa relação não é linear. De fato, a CCI tem a forma de "s" com inclinação e deslocamento na escala de habilidade definidos pelos parâmetros do item.

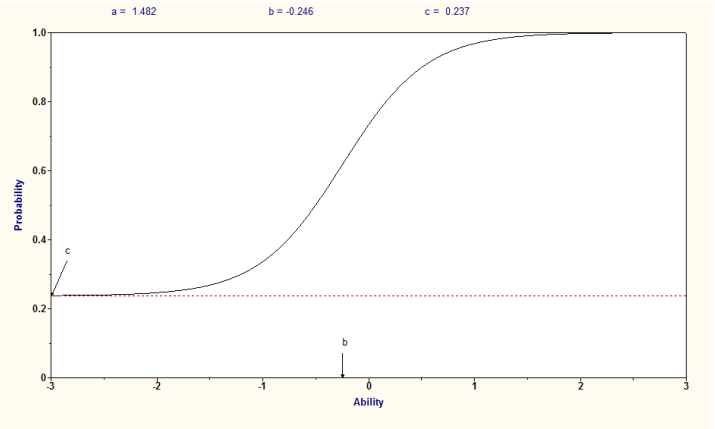



Figura 1 - Curva Característica do Item (CCI).
Fonte: Souza, 2016, p. 96

Definidos os itens do teste e conhecidas as suas curvas, a proficiência de cada respondente é obtida considerando-se essas curvas e as respostas escolhidas pelo aluno para cada item. A proficiência atribuída a cada aluno é o valor que dá maior probabilidade de observação do perfil de respostas produzido por cada aluno. Isso é uma aplicação do princípio da máxima verossimilhança, no qual estão baseadas todas as técnicas estatísticas de análise de dados. Como resultado, a proficiência do aluno não é equivalente ao número de itens acertados pelo aluno, mas de quais itens foram acertados pelos alunos. A metodologia busca coerência das respostas dos alunos e essa é a mudança de maior visibilidade em relação ao processo usual.

As utilizações de itens âncoras, ou itens de ligação, fazem com que seja possível comparar os resultados de uma prova com outras, ou mesmo aplicar provas diferentes para grupos distintos. Para a comparação dos resultados de forma mais direta, constrói-se uma escala de proficiência, a proficiência estimada pelo modelo estatístico, que a TRI torna possível através da atribuição de escores e da análise. Uma base numérica de comparação, colocando os indivíduos mesmo de provas diferentes sobre a mesma escala (RABELO, 2013).

Os parâmetros dos itens e das habilidades são obtidos por cálculos matemáticos sofisticados, que somente podem ser feitos com o uso de computador, tomando-se como base uma matriz de respostas que traz em suas colunas as respostas dos itens que compõem o teste e cada uma de suas linhas refere-se a um indivíduo que respondeu ao teste. Esses cálculos são feitos por softwares específicos que utilizam de técnicas estatísticas e matemáticas de estimação.

Para uma análise qualitativa da escala numérica, na TRI, faz-se necessário um processo denominado de

interpretação pedagógica das classes de proficiência. O processo consiste na identificação de itens âncoras, representativos de cada nível da escala, a partir dos dados psicométricos e na interpretação pedagógica feita por especialistas após a aplicação do teste relativa ao real significado da avaliação proposta em cada item. Com esse processo, localizam-se assim os itens na escala de proficiência assim como se faz com os indivíduos. Esse é um dos grandes ganhos da TRI, que permite o estabelecimento de um feedback qualitativo para os estudantes, interpretando-se o significado da proficiência numérica em termos de aprendizado. (RABELO, 2013).

## 2. Dos pressupostos da Teoria de Resposta ao Item

A TRI impõe alguns pressupostos na construção dos itens, como a unidimensionalidade e a independência local, ou seja, cada item deve trabalhar uma aptidão de forma dominante e não ser dependente da resposta de outro item (RABELO, 2013). Logo, a probabilidade de acerto de um item depende exclusivamente do traço medido e do comportamento do item e não da ordem do item dentro do teste, das condições físicas do examinando no momento do teste etc., e, desse modo, cada item informa algo sobre a proficiência do aluno. Outro pressuposto é da monotonicidade, ou seja, a probabilidade da resposta correta a um item deve aumentar à medida que aumenta o desempenho dos indivíduos no teste como um todo. Quando a TRI se classifica como robusta, tem correlação com a mínima violação desses três pressupostos (COHEN; SWERDLIK; STURMAN, 2014). Esses pressupostos não podem ser empiricamente demonstrados ou possuem bases lógicas, mas são provadas indiretamente, isto é, verificando se a sua violação produz resultados contraditórios em casos práticos (PASQUALI, 2003).

### 2.1 Unidimensionalidade

No que se refere à unidimensionalidade, o item deve avaliar apenas um traço latente, ou seja, espera-se que haja apenas um construto responsável pela realização do item. Há controvérsias em relação à unidimensionalidade dos testes, pois esse pressuposto nunca pode ser plenamente satisfeito uma vez que fatores cognitivos, de personalidade e de testagem, podem afetar o desempenho do submetido ao teste. Habilidades para responder ou não rapidamente, ansiedade, motivação, etc., são fatores que afetam a unidimensionalidade do teste (ANDRADE; LAROS; GOUVEIA, 2010). Assim, para satisfazer a unidimensionalidade, basta que se tenha um fator dominante responsável pelas respostas dos avaliados.

McDonald (1981) define a dimensionalidade baseada nas teorias de traço latente. Essa abordagem assume que k traços latentes explicam a realização dos sujeitos, de tal forma que, para um nível fixo de capacidade, as respostas dos sujeitos aos itens seriam estatisticamente independentes (VITÓRIA; ALMEIDA; PRIMI, 2006). Quando um único traço explica a realização dos sujeitos, então esse conjunto de itens considera-se unidimensional (HAMBLETON; SWAMINATHAM, 1985).

Ao se considerar o conceito de dimensionalidade no contexto de TRI, quando se diz que um teste é unidimensional o que se quer dizer é que todos os sujeitos que possuem a mesma capacidade estimada têm a mesma probabilidade de dar uma resposta correta a cada item (CUESTA, 1996). Essa definição não pressupõe que a capacidade seja função de um único traço, basta que a combinação de traços seja idêntica para todos os itens.

Vitória, Almeida e Primi apontam como dificuldade a determinação da unidimensionalidade não ter critérios empíricos consensuais, sendo assumida como uma questão de grau, ou seja, a resposta para a questão de qual o grau, ou qual o ponto de corte acima do qual podemos considerar a unidimensionalidade, não foi completamente respondida, deixando algum espaço à sensibilidade e bom senso (VITÓRIA; ALMEIDA; PRIMI, 2006). Pasquali (2003) considera a unidimensionalidade como uma questão de grau, pois o desempenho humano é multimotivado e multideterminado e, assim, em qualquer ação de um respondente ao teste, há a tendência que mais de um traço latente esteja presente. Hambleton e Swaminathan (1985) indicam que a unidimensionalidade não pode ser estritamente atingida. Assim, não se trata de unidimensionalidade "pura" de um item, mas em que medida as dimensões extras geram distorção na medida do construto.

No caso particular da SEEDF, nos últimos anos, o SIPAEDF se vale de itens calibrados e estruturados em boas práticas de elaboração nos testes procedimentais já realizados, diminuindo a influência do formato dos itens e do caderno de provas na medida dos traços latentes. Além disso, há a cultura de aplicação de avaliações externas, provas diagnósticas e simulados, tanto via SIPAEDF quanto via Saeb (Nacional) ou Pisa (internacional), de modo que professores e estudantes da rede pública e privada do Distrito Federal já estão habituados a participar de avaliações externas. Assim, busca-se o controle de fatores adicionais, tais como, desmotivação, ansiedade ou marcação errada em folha de respostas, de modo a tornar a medida de

desempenho ainda mais realista. As práticas recentes da DIAV para construção e aplicação das provas apontam, deste modo, para a garantia de níveis interessantes de unidimensionalidade dos itens na Prova DF.

## 2.2 Independência local

A independência local diz respeito ao fato de que as respostas dos respondentes aos itens são estatisticamente independentes se mantidas constantes as aptidões que afetam o teste, menos a aptidão dominante. De outro modo, o desempenho do avaliado em um item não afeta o seu desempenho nos demais itens.

O conceito de unidimensionalidade aparece no campo conceitual muito associado à independência local e, em alguns momentos, são tomados como sinônimos. Apesar de o conceito de independência local ser imprescindível para a definição de traço latente e de dimensionalidade, não podem ser tomados como sinônimos, pois a independência local pode ser conseguida com n dimensões (VITÓRIA; ALMEIDA; PRIMI, 2006).

A independência local implica que a sequência de respostas de um sujeito a uma série de itens será o produto das probabilidades de cada item individual.

Pasquali (2003) distingue que, por mais provável que as respostas de um mesmo respondente sejam correlacionadas, a independência local indica que, se houver correlação, esta se deve a fatores estranhos e não à aptidão que está sendo avaliada. Se as dimensões externas forem controladas, o fator dominante será a única fonte de variação e, assim, as respostas se tornam independentes pelo fato de o respondente responder o item em função do traço latente avaliado. Em geral, quando o pressuposto da unidimensionalidade é satisfeito, o pressuposto da independência local também o é (ANDRADE; LAROS; GOUVEIA, 2010).

Não há independência local quando a informação de um item ajuda a responder outro item ou quando a informação para a resposta correta é dada dentro do item, pois a habilidade para detectar a informação é uma dimensão além da habilidade sendo testada. A melhor maneira de lidar com a dependência local é prevenindo a sua ocorrência (EMBRESTON; REISE, 2000). Deste modo, o instrumento avaliativo deve ser constituído de itens elaborados considerando a engenharia de itens, que Rabelo (2013) define como boas práticas para elaboração de itens, de modo que a habilidade a ser avaliada seja de fato alcançada no item criado com essa finalidade. No caso particular do SIPAEDF, A Portaria nº 38, de 18 de fevereiro de 2020 (DISTRITO FEDERAL, 2020), prevê a criação de um banco distrital de itens por parte da SEEDF, que começou a formar elaboradores de itens em 2020 em formação coordenada pela SUPLAV e pela Subsecretaria de Formação Continuada dos Profissionais de Educação (EAPE) com o objetivo de criar itens na perspectiva da independência local.

## 2.3 Monotonicidade

Quando se trata de testes educacionais, é razoável supor que um sujeito com maior aptidão, isto é, que possui um nível mais elevado no processo do construto que determinado item mede, terá uma probabilidade de maior de acertar este item do que um sujeito com nível inferior de aptidão. Como na TRI a proficiência é descrita em termos de probabilidade, suponha que t seja a proficiência, então a probabilidade de acerto é definida por pi(t), ou seja, a probabilidade p de acertar o item i dado um tamanho de proficiência t. Deste modo, um indivíduo de menor proficiência terá uma pi(t) pequena enquanto um de maior proficiência terá pi(t) maior. Por se tratar de probabilidade, o valor de pi(t) varia de 0 a 1, onde 0 indica o indivíduo com nenhuma aptidão e 1 indica o indivíduo com aptidão ótima. Essa relação não é linear, mas é representada na CCI, conforme o gráfico da figura 1.

# 3. Aplicações da TRI nas avaliações de larga escala

Atendidos os seus pressupostos e com os dados ajustados ao modelo, a Teoria de Resposta ao Item permite conclusões interessantes e bem fundamentadas sobre o desempenho de uma população ao longo do tempo, ao inserir os desempenhos dos indivíduos em uma mesma escala de medida. A TRI viabiliza a comparação de resultados de indivíduos avaliados em épocas diferentes com a construção de escalas de proficiência. Essa propriedade é de grande importância para gestores da área de educação, pois permite acompanhar a evolução do sistema de ensino e embasa a tomada de decisão para as políticas públicas relacionadas ao tema.

Como caso nacional de sucesso pode-se citar o Saeb, seguramente um dos mais importantes sistemas de avaliação educacional do país pelo alcance e relevância, que é precursor do uso da TRI no Brasil na área de avaliação em larga escala. A TRI é utilizada nos processos relacionados ao Saeb desde 1995 e desde então permite o acompanhamento da evolução do sistema educacional brasileiro ao longo dos anos.

O caso Enem 2010, citado na introdução deste trabalho, revela outra propriedade importante da TRI, que é permitir que vários testes sejam aplicados no mesmo ano sem abrir mão da isonomia entre seus participantes. Outro teste, agora de abrangência e consolidação internacional, é Test of English as a Foreign Language

(TOEFL), exame de aptidão de língua inglesa aceito por mais de 11 mil universidades e instituições em mais de 150 países, que é aplicado várias vezes no ano e cada examinando recebe uma prova diferente, podendo realizá-la em algum centro de aplicação ou em sua própria residência, com a garantia de todos os resultados serem comparáveis e considerados isonômicos .

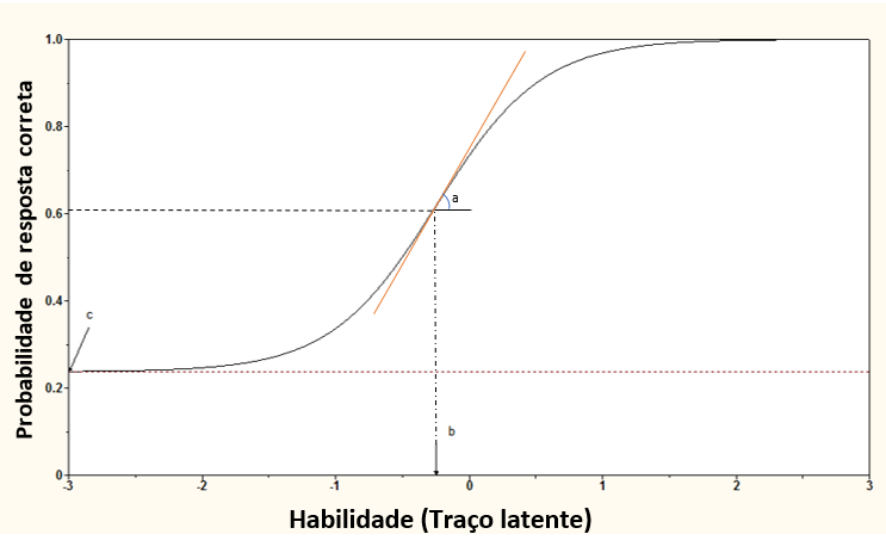
Para que essas propriedades sejam satisfeitas, o teste precisa seguir alguns protocolos na sua construção e na análise de seus resultados. As habilidades a serem avaliadas são organizadas em uma matriz de referência . Para atender os pressupostos de unidimensionalidade e de independência local, o elaborador de item seleciona uma única habilidade na matriz de referência e constrói o item considerando as práticas da engenharia de itens. A partir do momento que há itens elaborados a ponto de suprir a matriz de referência, eles passam por calibragem, processo para verificar a adequação de cada item para atingir o construto que se pretende avaliar. Com os itens calibrados, monta-se o instrumento avaliativo, o caderno de provas, com a quantidade de questões que abrange a matriz de referência levando-se em consideração o equilíbrio ou proporção do nível de dificuldade de cada item. Após a aplicação dos cadernos de prova e gerada a proficiência de cada examinando, o desempenho é descrito numa escala de proficiência.

Neste trabalho, foi centralizado o uso de itens do tipo múltipla escolha, que são amplamente utilizados no Saeb e no Enem, que a Portaria nº 38, de 18 de fevereiro de 2020 (DISTRITO FEDERAL, 2020), indica a ser utilizado no SIPAEDF, e ainda aproveitado em outros sistemas de avaliação, como o SARESP .

### 3.1 Sobre parâmetros, calibragem de itens e montagem dos cadernos de prova

Modelo da TRI mais utilizado nas avaliações de larga escala é o modelo logístico de três parâmetros, que relaciona a proficiência θ do respondente com os seguintes parâmetros: a discriminação a, a dificuldade b e o acerto casual c. A discriminação é definida como o poder do item para diferenciar indivíduos com magnitudes próximas da habilidade que está sendo aferida (RABELO, 2013). A dificuldade do item na TRI é o nível mínimo de proficiência que o estudante precisa possuir para ter uma chance alta de acertar a resposta. O índice de acerto casual representa a probabilidade de um indivíduo com baixa habilidade responder corretamente ao item.

Figura 2 - CCI com indicação dos parâmetros
Fonte: elaboração própria




Em termos de CCI, o parâmetro a é proporcional à derivada da tangente da curva característica no ponto de inflexão. Assim, nesse modelo, não se espera itens com discriminação negativa, pois indicaria que a probabilidade de responder corretamente ao item diminui com o aumento da habilidade, contrariando o pressuposto de monotonicidade. Um valor baixo para a discriminação significa que o item tem baixo poder de discriminação, ou seja, indivíduos com níveis muito distantes de proficiência apresentam probabilidades próximas de responder corretamente ao item. Um valor alto para a discriminação, por sua vez, implica na curva característica mais íngreme e separam bem examinandos que possuem habilidade abaixo do valor de b dos que possuem habilidade acima do valor do parâmetro b. De modo geral, itens com CCI mais inclinadas são mais úteis para diferenciar indivíduos com habilidades diferentes do que itens com a CCI mais achatada.

O parâmetro b representa a habilidade necessária para uma probabilidade de acerto igual a ((1+c))/2, logo, quanto maior o valor de b, mais difícil é o item.

O parâmetro c representa a probabilidade de um indivíduo com baixa habilidade responder corretamente o item, por isso é chamado de probabilidade de acerto casual, ou de probabilidade de acerto ao acaso, conhecido popularmente como "chute". Na CCI, o parâmetro de acerto casual se manifesta pela assíntota da curva. Como se trata de uma probabilidade, pode

variar entre zero e um e é desejável que o seu valor seja menor que ou igual a 0,2 (caso o item de múltipla escolha tenha cinco alternativas) ou que 0,25 (no caso do item de múltipla escolha tenha quatro alternativas), valores que equivalem à probabilidade de se acertar a resposta do item a partir de uma marcação aleatória.

Como já foi visto, os modelos da TRI representam a probabilidade de um indivíduo dar uma certa resposta a um item como função dos parâmetros do item e da habilidade do examinando. Em geral, não se conhece nem os parâmetros dos itens e nem a habilidade do respondente, havendo apenas a resposta deste ao teste. Daí, o processo de estimação dos parâmetros dos itens e das habilidades dos examinandos serem processos de grande relevância.

Calibragem pode ser definida como o processo de estimação de parâmetros dos itens. Para realizar a calibragem dos itens, é necessário aplicar em um grupo que seja significativo, ou seja, que tenha uma quantidade de examinandos suficiente para gerar dados robustos para a TRI. Não se trata de uma decisão trivial: se por um lado esse grupo não pode ser muito grande (o custo pode tornar-se muito alto o que poderia inviabilizar a criação do caderno de provas; muitos indivíduos tendo acesso aos itens passariam a comprometer o sigilo do teste, etc.) por outro lado não pode ser muito pequeno, sob pena de não ser representativa e, assim, não estimar adequadamente os parâmetros psicométricos. Estudos apontam que os parâmetros dos itens podem ser estimados adequadamente para amostras a partir de duzentos participantes (NUNES; PRIMI, 2005; HAMBLETON; SWAMINATHAM, 1985; MUÑIZ, 1990).

Os parâmetros dos itens e das habilidades são obtidos por sofisticados cálculos matemáticos, que somente podem ser feitos com o uso de recursos de informática que utilizam técnicas estatísticas e matemáticas de estimação por meio de softwares específicos, tomando-se por base uma matriz de respostas que traz em suas colunas as respostas dos itens que compõem o teste e cada uma de suas linhas refere-se a um indivíduo que respondeu ao teste (SOUZA, 2016). Andrade, Tavares e Valle (2000) apontam e descrevem o Método da Máxima Verossimilhança como metodologia mais comum para a determinação dos parâmetros e ressaltam a importância do avanço do campo da informática. Andrade, Laros e Gouveia (2010) sugerem o uso do software Bilog-MG 3.0 para realizar a estimação e descreve as suas funções para realizar a calibragem de itens. Ainda é possível utilizar outros softwares para estimar os parâmetros, como o R, amplamente utilizado em pesquisa envolvendo estatísticas, mas que, ao contrário do Bilog-MG 3.0, não exige licença para o seu uso.

É importante frisar que o uso da TRI não exclui o uso da TCT no processo de calibragem. Ao contrário do que se pode pensar a respeito, a TCT traz informações preciosas a respeito da qualidade dos itens e por isso os seus parâmetros também devem ser analisados no processo. A discriminação pela TCT pode indicar itens a serem descartados ou com necessidade de reelaboração e os coeficientes bisseriais das alternativas podem acusar atratividade de indivíduos com alta habilidade por parte de algum distrator (alternativa inequivocamente errada), o que é conhecido popularmente como "peguinha" (SOUZA, 2016).

A partir do momento em que há itens calibrados a ponto de cobrir toda a matriz de referência, é possível montar o teste. A Tabela 1 representa a distribuição e a classificação de acordo com a dificuldade dos itens recomendada por Rabelo (2013) e adotada por diversos autores da área de avaliação e psicometria.

Tabela 1 - Classificação e proporção de itens no Teste de acordo com o parâmetro dificuldade.

| Valor de b | Classificação | % esperada |
|---|---|---|
| até -1,28 | muito fácil | 10 |
| -1,27 ≤ b ≤ -0,52 | fácil | 20 |
| -0,51 ≤ b ≤ 0,51 | mediano | 40 |
| 0,52 ≤ b ≤ 1,27 | difícil | 20 |
| 1,28 ou mais | muito difícil | 10 |

Fonte: Rabelo, 2013, p. 134

### 3.2 Escala de proficiência

Escala de proficiência é um instrumento de estimação das habilidades dos respondentes. A proficiência é estimada pelo modelo estatístico, que a TRI torna possível através da atribuição de escores e da análise, de

modo a colocar os indivíduos numa mesma escala mesmo que tenham realizado provas diferentes.

Para a construção da escala de proficiência é necessário escolher uma origem, que é estabelecida no valor médio, e uma escala de medida, o desvio padrão, ambos obtidos das proficiências dos indivíduos que responderam ao teste. Mantém-se uma relação de ordem entre os valores da escala, quando se comparam duas escalas diferentes, sempre relacionados com o valor médio e com o desvio padrão.

É importante ressaltar que independente da escala com a qual se está trabalhando, a probabilidade de um indivíduo responder corretamente a certo item é sempre a mesma, ou seja, a proficiência de um indivíduo é invariante em relação à escala de medidas. Portanto, independentemente da escala adotada, os resultados serão os mesmos e a interpretação feita sob o olhar das duas escalas é a mesma. Por exemplo, a escala (500,100), valor médio igual a 500 e desvio padrão igual a 100, é utilizada no Enem. Um indivíduo com habilidade 2 na escala (0,1) tem proficiência de 2 desvios-padrão acima da média, correspondendo na escala (500,100) à habilidade 700, pois também representa 2 desvios-padrão acima da média.

A elaboração da escala de proficiência passa pelo processo de equalização. Em relação à TRI, equalizar significa colocar parâmetros de itens vindos de provas distintas ou habilidades de respondentes de diferentes grupos, na mesma métrica, isto é, numa escala comum, tornando os itens e/ou as habilidades comparáveis (ANDRADE; TAVARES; VALLE, 2000).

Uma vez que todos os parâmetros dos itens e que todas as habilidades dos respondentes ⎡ tanto individuais como populacionais ⎡ de todos os grupos avaliados estão numa mesma métrica, ou seja, quando todos os parâmetros envolvidos são comparáveis, pode-se então construir escalas de conhecimento interpretáveis. As proficiências de todos os alunos que participam de uma avaliação externa normalmente são organizadas em uma escala, a Escala de Habilidade ou Proficiência.

Devido à natureza arbitrária das estimativas dos parâmetros dos itens e das habilidades, é possível comparar entre si habilidades obtidas para diferentes respondentes, mas que não possuem, por si mesmas, significado pedagógico. Esse fato motivou então a criação de escalas de habilidades ou de proficiência.

Uma escala só é útil para finalidade de diagnóstico ou de ação pedagógica se os seus diferentes pontos tiverem uma interpretação pedagógica. A interpretação começa com a construção de um mapa de itens, que consiste em associar cada item do teste a um ponto da escala utilizada para medir proficiências dos alunos. A relação de um item com a escala é probabilística e, frequentemente, toma-se como ponto de locação a proficiência em que a probabilidade de acertar o item seja de 65%. Construído o mapa de itens, a interpretação pedagógica considera que alunos com proficiência em um dado valor são capazes de fazer tarefas implícitas nos itens localizados em valores menores. Existem várias nuances para aplicação desse princípio e a referência clássica é o artigo de Beaton e Allen (1992).

Para que seja possível construir uma escala de proficiência interpretável é conveniente que se tenha uma matriz de referência, na qual estejam estabelecidas competências (ou descritores) que abordam as habilidades que se deseja medir. O instrumento de medida para o traço latente investigado costuma ser uma prova composta de itens elaborados de acordo com as habilidades que se deseja aferir, dispostas na matriz de referência.

A escala é produzida, primeiramente, com dados brutos dos desempenhos dos alunos e seu entendimento, em geral, é mais difícil para o público em geral. Então, com o intuito de facilitar a compreensão, é feita uma interpretação da escala, e os dados são apresentados através de referências comuns, inteligíveis ao público.

A interpretação da escala é realizada mediante o processo de incorporação de informações normativas e pedagógicas, o que permite evidenciar os conhecimentos dos alunos.

As informações normativas descrevem e revelam como um aluno está em um momento específico do seu processo de aprendizagem. Incorporar significado normativo à escala é uma maneira de aumentar sua interpretabilidade através de conceitos, ou do uso de etiquetas para certos níveis ou categorias (SOARES, 2009). De maneira geral, a escala de zero a 10 permite julgamentos apoiados no senso comum: se o aluno obtém 8 pontos num teste que vale 10 pontos é usual acreditar que ele teve bom desempenho. Da mesma maneira, as pessoas conseguem lidar e entender conceitos do tipo A, B, C e D ou muito bom (MB), bom (B) e regular (R), porque estes fazem parte de uma estrutura normativa.

Nas avaliações externas, as informações normativas aparecem nas etiquetas que são atreladas aos níveis de desempenho. Por exemplo, o Índice de Desenvolvimento da Educação do Estado de São Paulo (Idesp) e o Saeb utilizam quatro níveis de desempenho: Abaixo do Básico, Básico, Adequado e Avançado, nos quais distribui seus avaliados (SOARES, 2009).

As escalas de proficiência tornam possível a interpretação pedagógica dos valores das habilidades. Essas escalas são definidas por níveis âncora, que por sua vez

são caracterizados por um conjunto de itens âncora. Níveis âncora são pontos selecionados pelo analista na escala de habilidade para serem interpretados pedagogicamente.

A construção da escala de proficiência decorre, em geral, do estabelecimento de pontos de corte e, por conseguinte, dos níveis de desempenho. Para isso, executam-se atividades para eleger itens cujos pontos de alocação definem os pontos de corte entre os níveis de desempenho. Para a realização dessa tarefa é necessário formar uma equipe capaz de desenvolvê-la, constituída tanto por especialistas das competências avaliadas, como por especialistas da área de Estatística. Cabem à parte pedagógica da equipe a determinação do número de níveis de desempenho, a seleção de seus nomes, a definição dos pontos de corte e a descrição pedagógica dos níveis. É responsabilidade dos estatísticos determinar, com a aplicação da TRI, o ponto de alocação dos itens na escala.

Quanto à quantidade de níveis, observa-se que, na prática, não são usados muitos, prevalecendo uma organização em torno de três a cinco níveis. Quatro parece um número ideal, porque assim se trabalha um nível negativo, que revela insuficiência de competência em um determinado domínio, e em três níveis positivos, desde aquele que reflete a competência mínima em um determinado domínio até aquele que expressa um grau de competência mais elevado. Como exemplo, o Saeb e o Saresp utilizam quatro níveis: Abaixo do Básico, Básico, Adequado e Avançado. O aluno classificado no nível Adequado demonstra dominar os conteúdos e habilidades esperados para o seu estágio escolar. Os do nível avançado dominam a competência de forma especialmente completa, ultrapassando o esperado para o seu estágio escolar. O nível Básico congrega os alunos que demonstram apenas o domínio apenas parcial da competência. Finalmente, os alunos de nível Abaixo do Básico mostram domínio rudimentar da competência medida. Na literatura, nomes alternativos, mas equivalentes, são frequentemente utilizados (SOARES, 2009).

## Considerações finais

Um teste educacional é um instrumento para realizar medida e, como tal, deve ser válido e preciso para que resultados emitidos sejam os mais próximos da realidade. No contexto de gestão pública na área de educação, são tomadas decisões que afetam a vida de milhares de pessoas e que geram custo muito alto para os cofres públicos. Caso o gestor público utilize resultados inválidos ou com pouca precisão, a verba deixa de ser investimento em educação para se tornar prejuízo ao erário e as políticas públicas operacionalizadas podem tomar rumos desastrosos.

Os conceitos apresentados nas seções do presente trabalho são úteis em um contexto de construção de instrumentos avaliativos e o profissional em contexto escolar precisa tê-las em mãos como ferramentas complementares ao seu trabalho de avaliação. A construção de uma prova com itens de múltipla escolha, com toda a reserva técnica que a torna válida, é uma tarefa difícil, porém possível e necessária para que o professor regente tenha subsídios para tomar as melhores decisões no planejamento de suas sequências didáticas. No Brasil, temos experiências bem sucedidas, tanto em nível federal quanto estadual e municipal, na construção e no uso de escalas de proficiência a partir da TRI e da TCT. Espera-se que, em especial, no âmbito da SEEDF haja aprofundamento nos estudos dos conceitos explanados nas seções deste texto, mas acima disso, que seja feita a aplicação na rede, principalmente em avaliação de larga escala, que resulte na melhor compreensão do nível de aprendizagem de seus estudantes e, consequentemente, na gestão eficaz das políticas públicas em educação no Distrito Federal.

Conclui-se, portanto, que a TRI se mostra como uma ferramenta importante para sistemas de avaliação de larga escala que desejam comparar o desempenho de seus estudantes ao longo do tempo e, assim, acompanhar o desenvolvimento e evolução da qualidade de ensino. Sem abrir mão dos elementos já bem estabelecidos da Psicometria Clássica, a TRI torna o processo mais justo e seus resultados mais fidedignos.

## Notas

1 Alguns estados optaram por não aplicar provas de avaliação em larga escala em 2020 devido ao contexto de pandemia, seguindo as orientações do Conselho de Educação (GONÇALVES, et al., 2020). A SEEDF não foi diferente, suspendendo a aplicação da Prova DF no ano em questão.

2 Disponível em: https://educacao.uol.com.br/noticias/2010/11/08/justica-determina-suspensao-do-enem-2010.htm. Acesso em: 30 jan. 2021.

3 Disponível em: https://educacao.uol.com.br/noticias/agencia-estado/2010/11/07/para-mpf-e-oab-erro-pode-anular-enem.htm. Acesso em: 30 jan. 2021.

4 O Programa Internacional de Avaliação de Estudantes (Pisa), tradução de Programme for International Student Assessment, é um estudo comparativo internacional realizado a cada três anos pela Organização para a Cooperação e Desenvolvimento Econômico (OCDE). O Pisa oferece informações sobre o desempenho dos estudantes na faixa etária dos 15 anos, idade em que se pressupõe o término da escolaridade básica obrigatória na maioria dos países, vinculando dados sobre seus backgrounds e suas atitudes em relação à aprendizagem, e também aos principais fatores que moldam sua aprendizagem, dentro e fora da escola. Disponível e: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/pisa Acesso em: 07 fev. 2021.

5 Disponível em: https://www.ets.org/pt/toefl/test-takers/ibt/why/accepted-preferred. Acesso em: 06

6 O termo matriz de referência é utilizado especificamente no contexto das avaliações em larga escala para indicar habilidades a serem avaliadas em cada etapa da escolarização e orientar a elaboração de itens de testes e provas, bem como a construção de escalas de proficiência que definem o que e o quanto o aluno realiza no contexto da avaliação. Trata-se de uma referência para a construção do instrumento de avaliação, sendo diferente de uma proposta curricular ou programa de ensino, que são mais amplos e completos. Disponível em: http://inep.gov.br/matrizes-de-referencia1 Acesso em: 07 fev. 2021.

7 O Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo (Saresp) é aplicado pela Secretaria da Educação do Estado de São Paulo com a finalidade de produzir um diagnóstico da situação da escolaridade básica paulista, visando orientar os gestores do ensino no monitoramento das políticas voltadas para a melhoria da qualidade educacional. No Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo (Saresp), os alunos do 3º, 5º, 7º e 9º anos do Ensino Fundamental e da 3ª série do Ensino Médio têm seus conhecimentos avaliados por meio de provas com questões de Língua Portuguesa, Matemática, Ciências Humanas, Ciências da Natureza e redação. Disponível em: https://saresp.fde.sp.gov.br/ Acesso em: 07 fev. 2021.

## Referências


ANDRADE, D. F.; DE, TAVARES, H. R.; VALLE R. C., Teoria de Resposta ao Item: conceitos e aplicações. ABE – Associação Brasileira de Estatística, São Paulo, 2000.

ANDRADE, J. M.; LAROS, J. A.; GOUVEIA, V. V. O uso da Teoria de Resposta ao Item em avaliações educacionais: diretrizes para pesquisadores. Aval. psicol., Porto Alegre, v. 9, n. 3, p. 421-435, dez. 2010.

ANDRIOLA, W. B., Psicometria Moderna: características e tendências, Est. Aval. Educ., São Paulo, v. 20, n. 43, maio/ago. 2009.

BEATON, A. E.; ALLEN, N. L. Interpreting scales through scale anchoring, Journal of Educacional Statistics, 17, 191-204, 1992.

COHEN, R. J.; SWERDLIK, M. E.; STURMAN, E. D. Testagem e avaliação psicológica: introdução a testes e medidas . 8. ed. Porto Alegre: AMGH, 2014.

CUESTA, M. Unidimensionalidade. In: J. Muñiz (Ed.), Psicometría. Madrid: Editorial Universitas. 1996.

DISTRITO FEDERAL. Portaria nº 38, de 18 de fevereiro de 2020. Normatiza o Sistema Permanente de Avaliação Educacional da Secretaria de Estado de Educação do Distrito Federal (SIPAEDF). Diário Oficial do Distrito Federal. 19 fev. 2020.

EMBRESTON, S. E.; REISE, S. P. Item response theory for psychologists. Lawrence Erlbaum. New Jersey, 2000.

GHEDIN, E.; FRANCO, M. A. S. F. Questões de método na construção da pesquisa em educação. 2. ed. São Paulo: Cortez, 2011. (Coleção docência em Formação. Série saberes pedagógicos).

GONÇALVES, L. F. A. et al. As políticas públicas de avaliação em larga escala no Brasil diante da pandemia de Covid-19. Revista Com Censo: Estudos Educacionais do Distrito Federal, [S.l.], v. 7, n. 3, p. 65-76, ago. 2020. ISSN 2359-2494. Disponível em: http://periodicos.se.df.gov.br/index.php/comcenso/article/view/932. Acesso em: 08 fev. 2021.

HAMBLETON, R. K.; SWAMINATHAN, H. Item Response Theory. Principles and Applications. Kluwer Nijhoff Publishing. Boston, MA. 1985.

MATHISON, S. Encyclopedia of Evaluation. Thousands Oaks. Sage Publications. 2005.

McDONALD, R. The dimensionality of testes and items. British Journal of Mathematical and Statistical Psychology, 34, 100-117. 1981.

MUÑIZ, J., Teoria de Respuesta a los Itens: un nuevo enfoque em la evolución psicológica y educativa. Madri: Ediciones Pirámide, SA, 1990.

MUÑIZ, J., Teoria Clássica dos Testes. Madri: Ediciones Pirámide, SA, 2003.

NUNES, C. H. S.; PRIMI, R. Impacto do Tamanho da Amostra na Calibração de Itens e Estimativa de Escores por Teoria de Resposta ao Item. Aval. psicol. Porto Alegre . v. 4, n. 2, p. 141-155, dez. 2005.

PASQUALI, L Psicometria: teoria dos testes na psicologia e na educação. Petrópolis-RJ: Vozes, 2003.

PASQUALI, L.; PRIMI, R. Fundamentos da Teoria de Resposta ao Item – TRI. Aval. psicol. Porto Alegre . v. 2, n. 2, p. 99-110, dez. 2003.

RABELO, M. L. Avaliação Educacional: fundamentos, metodologia e aplicações no contexto brasileiro. Rio de Janeiro: SBM, 2013.

SILVEIRA, D. T.; CÓRDOVA, F. P. A pesquisa científica. In: GERHARDT; T. E.; SILVEIRA, D. T. Métodos de pesquisa. Porto Alegre: Editora UFRGS, 2009. p. 31-42. Disponível em: MET.PESQUISA.indd (cesadufs.com.br) Acesso em: 12 dez. 2020.

SOARES, J.F. Índice de desenvolvimento da Educação de São Paulo – Idesp: bases metodológicas. São Paulo em Perspectiva. São Paulo, Fundação Seade, v. 23, n. 1, p. 29-41, jan./jun. 2009.

SOUZA, J. S. Projeto Simulado de uma escola pública do DF: convergência para uma avaliação formativa em Matemática. Dissertação de mestrado. UnB. Brasília, 2016.

VITÓRIA, F; ALMEIDA, L. S.; PRIMI, R. Unidimensionalidade em testes psicológicos: conceito, estratégias e dificuldades na sua avaliação. PSIC - Revista de Psicologia da Vetor Editora, v. 7, nº 1, p. 1-7, Jan./Jun. 2006.