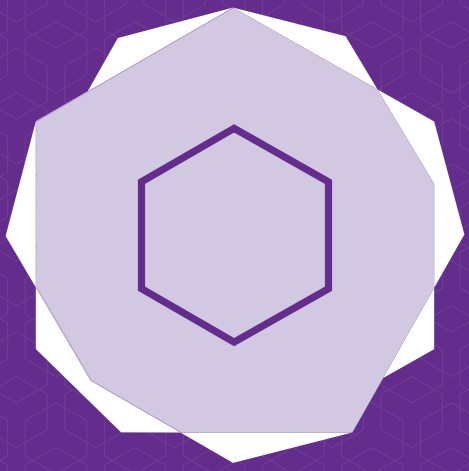


EXAMEN

Pesquisa em Avaliação, Certificação e Seleção

Cebraspe

N. 1

V.1 • jul.- dez. • 2017
ISSN 2526 - 9259

EXAMEN

Pesquisa em Avaliação, Certificação e Seleção

Revista semestral, volume 1, número 1, julho – dezembro 2017

ISSN 2526-9259

Cebraspe

Brasília, DF — Brasil

Cebraspe

Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos (Cebraspe), pessoa jurídica de direito privado, sem fins lucrativos, qualificado por meio do Decreto nº 8.078/2013 como Organização Social (OS), supervisionado pelo Ministério da Educação (MEC), mediante contrato de gestão, com a interveniência da Fundação Universidade de Brasília (FUB) e do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), tendo como finalidade precípua fomentar e promover o ensino, a pesquisa científica, o desenvolvimento tecnológico, o desenvolvimento institucional e a difusão de informações, experiências e projetos de interesse social e utilidade pública nas áreas de avaliação, certificação e seleção.

Conselho de Administração

Órgão/entidade	Conselheiro(a)	
Ministério da Educação	Vicente de Paula Almeida Júnior Weber Gomes de Sousa	Titular Suplente
Ministério da Ciência, Tecnologia, Inovações e Comunicações	Luiz Fernando Fauth Caroline Menicucci Salgado	Titular Suplente
Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep)	– Célia Cristina de Souza Gedeon Araújo	Titular Suplente
Rede Nacional de Ensino e Pesquisa (RNP)	Nelson Simões da Silva Wilson Biancardi Coury	Titular Suplente
Associação Brasileira de Estatística	Antonio Eduardo Gomes Hélio dos Santos Migon	Titular Suplente
Universidade de Brasília (UnB)	– Leonardo Rodrigues Araújo Xavier de Menezes	Titular Suplente
Associação dos Aposentados da FUB	Hildebrando de Miranda Flor Luiz Hernan Rodrigues Castro	Titular Suplente
Universidade de Brasília – Conselho Universitário	Alessandro Borges de Sousa Oliveira Marilde Loiola de Menezes	Titular Suplente
Universidade de Brasília – Conselho de Administração	Rodrigo Rosal Cavalcanti Santos Francisco de Assis Rocha Neves	Titular Suplente
Universidade de Brasília – Conselho de Ensino, Pesquisa e Extensão	Jurandir Rodrigues de Souza Alexandre Ricardo Soares Romariz	Titular Suplente
Representante dos Associados Fundadores	Noraí Romeu Rocco Marcelo Ladeira	Titular Suplente

Diretor-Geral
Paulo Henrique Portela de Carvalho

Diretora Executiva
Maria Osmarina do Espírito Santo Oliveira

Diretor de Contratação e Gestão de Eventos
Ricardo Bastos Cunha

Diretor de Instrumentos de Avaliação, Seleção e Certificação
Marcus Vinícius Araújo Soares

Diretor de Operações em Eventos
Jorge Amorim Vaz

Conselho Editorial do Cebraspe
José Otávio Nogueira Guimarães
Rogério Basali
Denise Aragão
Lucília Garcez
Mauro Luiz Rabelo

Examen

Examen publica artigos, resenhas e entrevistas que discutem avaliação educacional em larga escala, políticas públicas em educação e certificação educacional e profissional. O propósito da revista é servir como fórum para a apresentação de pesquisas atuais e como veículo de disseminação de informação para comunidade acadêmica, profissionais e sociedade em geral. O periódico é fomentado pelo Cebraspe e é publicado semestralmente.

Artigos, entrevistas e resenhas da publicação não refletem a opinião do Cebraspe ou da revista.

Examen : pesquisa em avaliação, certificação e seleção / Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos – Ano 1, n. 1 (jul. 2017).
Semestral

ISSN 2526-9259

1. Educação – Periódico. I. Brasil. 2. Educação – Políticas Públicas. 3. Avaliação Educacional.

CDU 371.26(81)

revistaexamen@cebraspe.org.br

SUMÁRIO

contents

APRESENTAÇÃO

contents

É com orgulho que apresentamos a Examen, primeira revista científica e de distribuição gratuita elaborada pelo Centro Brasileiro de Pesquisa em Avaliação e Seleção e de Promoção de Eventos (Cebraspe), organização social sem fins lucrativos vinculada ao Ministério da Educação (MEC) com a interveniência do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) e da Fundação Universidade de Brasília (FUB). A criação do centro foi aprovada pelo Conselho Universitário da UnB e configura-se, portanto, um projeto acadêmico.

A Examen publica artigos, resenhas e entrevistas que discutem avaliação educacional em larga escala e políticas públicas em educação, certificação e seleção de pessoas. Tem como objetivo divulgar a produção científica e os conhecimentos decorrentes dos programas de ensino e pesquisa em avaliação, certificação e seleção do Cebraspe; estimular a redação de artigos que reflitam sobre relatórios e estudos analíticos realizados com base em experiências e dados extraídos desses eventos; e promover a interação entre especialistas na área de educação.

A revista colabora, assim, para o alcance de dois importantes objetivos estratégicos da organização: produzir e disseminar conhecimentos na área de avaliação, certificação e seleção; e contribuir para a definição de políticas públicas em educação, por meio da publicação de textos, artigos, resenhas, entrevistas e documentos relativos à produção de conhecimento dessa área.

Desse modo, intitulamos a revista – que debate avaliações em larga escala e políticas públicas educacionais – Examen, termo de origem latina que apresenta, entre outras, duas acepções: ação de verificação, exame; e multidão, grande número de pessoas.

Esperamos que sua experiência com a revista Examen seja enriquecedora.

Boa leitura!

Prof. Paulo Henrique Portela de Carvalho

EDITORIAL

editorial

Temos o prazer de apresentar aos leitores o primeiro número da revista *Examen*, um novo periódico científico voltado para a disseminação de pesquisas relacionadas às temáticas de avaliação educacional em larga escala, políticas públicas em educação e certificação educacional e profissional. Trata-se de um periódico aberto, cujo conteúdo está disponível livremente. Os trabalhos serão publicados semestralmente e divulgados nas formas impressa e eletrônica. Neste exemplar, são oferecidos cinco estudos que abordam diversos temas relacionados a avaliação educacional, uma resenha sobre a terceira edição do livro do professor Thomas M. Haladyna: *Developing and validating multiple-choice test items,* além de uma entrevista com o professor Jacob Arie Laros, sobre o tema Psicometria.

Na seção de artigos, iniciamos com o trabalho de autoria do renomado professor Luiz Pasquali, cujo título é: "Validade dos testes". Neste estudo, o objetivo principal é apresentar o conceito de validade, as principais formas de medi-la e a história dessa relevante característica dos testes educacionais.

É fundamental que o conteúdo dos testes esteja alinhado com o currículo que se pretende avaliar. Sobre esse tema escreveram os autores Leah T. Kaira e Stephen G. Sireci no segundo artigo, "Using item mapping to evaluate alignment between curriculum and assessment", em que apresentam uma forma de avaliar o alinhamento do teste ao currículo com base no nível de dificuldade esperado.

"Proposta de segmentação de uma escala da TRI utilizando o nível socioeconômico" é o terceiro artigo, desenvolvido por Gabriela Thamara de Freitas Barros, Adriano Ferreti Borgatto e Adolfo Samuel de Oliveira. Essa pesquisa teve como finalidade propor uma alternativa para a construção e a interpretação de uma escala para um indicador de nível socioeconômico, utilizando-se o modelo politômico de respostas graduais da Teoria de Resposta ao Item (TRI).

"Estudos brasileiros sobre a eficácia escolar: uma revisão de literatura" é o quarto artigo, o qual traz uma pesquisa realizada por Camila Akemi Karino e Jacob Arie Laros com o objetivo de realizar uma revisão sistemática da literatura brasileira na área de eficácia escolar.

Por fim, o último artigo: “A avaliação no Plano Nacional de Educação – PNE (2014–2024)”, de autoria de Catarina de Almeida Santos e Danielle Xabregas Pamplona Nogueira, traz uma análise do tema avaliação no PNE, buscando compreender a concepção de avaliação presente no plano e sua relação com o conceito de educação, além de trazer apontamentos quanto ao Sistema Nacional de Avaliação da Educação Básica (Sinaeb).

Na segunda seção da revista, temos uma resenha do livro do professor Thomas M. Haladyna: *Developing and validating multiple-choice test items,* intitulada “Construindo testes: como elaborar e validar itens de múltipla escolha”.

A revista finaliza com a entrevista "Psicometria em avaliações", que apresenta o ponto de vista do professor Jacob Arie Laros – Ph.D. pela Universidade de Groningen, na Holanda, e atualmente professor titular no Instituto de Psicologia da Universidade de Brasília – a respeito da contribuição da Psicometria para a avaliação educacional.

Desejamos a todos uma proveitosa leitura!

Girlene Ribeiro de Jesus
Joaquim José Soares Neto
Dalton Francisco Andrade

ARTIGOS

articles

VALIDADE DOS TESTES

TEST VALIDITY

VALIDEZ DE LOS TESTES

Luiz Pasquali

RESUMO

A validade ocupa uma posição central na teoria da medida, constituindo-se um parâmetro fundamental e indispensável. Atualmente, é definida como a medida em que as evidências empíricas embasam as interpretações e os usos propostos para o teste. Neste estudo, o objetivo principal é apresentar o conceito de validade, a história desse parâmetro e as principais formas de medi-lo. A primeira parte do artigo explora as bases conceituais da história do parâmetro em três períodos. Em cada um deles, há a predominância de um dos tipos atualmente conhecidos de validade. Em seguida, são detalhados os procedimentos qualitativos e quantitativos recomendados para investigar validade na visão atual. Por fim, é apresentado o conceito de validade ecológica, que não constitui uma nova forma de coletar evidências de validade, mas sim de identificar como tais evidências devem ser buscadas.

Palavras-chave: validade; medidas educacionais; psicometria.

ABSTRACT

Validity occupies a key position in measurement theory, constituting a fundamental and indispensable parameter. Currently, it is defined as the degree to which empirical evidence supports interpretation and proposed test uses. The primary objective of this study is to present the concept of validity, the history behind this parameter, and the main ways of measuring it. The first part of the article explores the conceptual roots of the history of the validity parameter in three periods; in each one of them, there is the predominance of one of the recognized types of validity over the others. Subsequently, we detail the recommended qualitative and quantitative procedures to investigate validity under the current paradigm. Finally, we present the concept of ecological validity, which should not be understood as a new form of collecting evidence validity, but rather as a form of identifying how such evidence is to be sought.

Keywords: validity; educational measures; psychometrics.

RESUMEN

La validez ocupa una posición central en la teoría de la medida, constituyéndose como un parámetro fundamental e indispensable. Actualmente, es definida como la medida en que las evidencias empíricas embazan las interpretaciones y los usos propuestos para el test. En este estudio, el objetivo principal es presentar el concepto de validez, la historia de ese parámetro y las principales formas de medirlo. La primera parte del artículo explora las bases conceptuales de la historia del parámetro de validez en tres períodos. En cada uno de ellos, la predominancia de uno de los tipos actualmente conocidos de validez. En seguida, son detallados los procedimientos cualitativos y cuantitativos que son recomendados para investigar la validez en la visión actual. Por fin, es presentado el concepto de validez ecológica, que no se constituye como una nueva forma de recoger evidencias de validez, pero sí identificar como tales evidencias deben ser buscadas.

Palabras clave: validez; medidas educacionales; psicometría.

Introdução

A validade constitui um parâmetro da medida tipicamente discutido no contexto das ciências psicossociais, que trabalham com a modelagem latente. Ela não é corrente em ciências físicas, por exemplo, embora haja nessas ciências ocasiões em que tal parâmetro se aplicaria. Nestas, a preocupação principal na medida se centra na questão da precisão, na dita calibração dos instrumentos. Essa é importante também na medida em ciências psicossociais, mas ela não tem nada a ver, conceitualmente, com a questão da validade. A razão disso está no fato de que a validade diz respeito ao aspecto da medida de ser congruente com a propriedade medida dos objetos e não com a exatidão com que a mensuração, que descreve essa propriedade do objeto, é feita. Em Física, o instrumento é um objeto físico que mede propriedades físicas; então parece fácil ver que a propriedade do objeto mensurante é ou não congruente com a propriedade do objeto medido. Tome, por exemplo, o caso da propriedade "comprimento" do objeto. O instrumento que mede essa propriedade, isto é, o metro, usa a sua propriedade de comprimento para medir o comprimento de outro objeto; então, mensura-se comprimento

com comprimento, tomados estes termos univocamente. Não há necessidade de provar que a propriedade "comprimento" do metro seja congruente com a mesma propriedade no objeto medido; os termos são unívocos, eles são conceitualmente equivalentes, aliás, idênticos.

O caso já se torna menos claro quando, por exemplo, o astrônomo mede a propriedade "velocidade" galáctica de aproximação ou afastamento via efeito Doppler, no qual a aproximação/afastamento das linhas espectrais da luz da galáxia seria o instrumento da medida. Aqui já temos, na verdade, um problema de validade do instrumento de medida, a saber, é verdade ou não que as distâncias das linhas espectrais têm a ver com a velocidade das galáxias? Pode-se fazer tal suposição, mas ela tem que ser demonstrada empiricamente de alguma maneira, isto é, pelo menos em suas consequências, em hipóteses dela derivadas ou deriváveis e verificáveis. Nesse caso específico, o problema da precisão da medida diz respeito a quão exata pode ser feita a mensuração das distâncias entre as linhas espectrais, ao passo que o problema da validade diz respeito ao fato de essa medida, por mais exata e perfeita que ela possa ser, ter algo a ver ou não com a velocidade de afastamento da galáxia. Em outras palavras, a validade em tal caso diz respeito à demonstração da legitimidade da representação ou da modelagem da velocidade galáctica via distâncias das linhas espectrais.

Esse caso da astronomia ilustra o que tipicamente acontece com a medida em ciências psicossociais e, consequentemente, torna a prova da validade dos instrumentos nessas ciências algo fundamental e crucial, isto é, é uma condição *sine qua non* demonstrar a validade dos instrumentos nessas ciências. Isso é particularmente o caso nos enfoques que, em Psicologia, trabalham com o conceito de traço latente, pelos quais se deve demonstrar a correspondência (congruência) entre traço latente e sua representação física (o comportamento). Não causa estranheza, portanto, que o problema de validade tenha tido, na história da Psicologia, uma posição central na teoria da medida, constituindo-se, na verdade, o seu parâmetro fundamental e indispensável. Aliás, a história desse parâmetro é repleta de diatribes que espelham concepções teóricas antagônicas da própria teoria psicológica. À questão de "como legitimar ou justificar a pertinência da medida do comportamento humano?" foram dadas respostas diferentes na história da Psicometria. Essa diatribe pode ser ilustrada distinguindo várias etapas

de predominância de uma concepção do parâmetro validade sobre outras e que aparecem sempre atreladas a uma concepção mais geral da própria Psicologia, como já anotava Anastasi em 1986.

Desenvolvimento

Com efeito, poder-se-ia delinear, em traços bem gerais, a história do parâmetro da validade em três períodos. Em cada um deles, há a predominância de um dos tipos atualmente conhecidos de validade, desde o famoso trabalho de Cronbach e Meehl (1955), expressos sob o modelo trinitário, a saber, validade de conteúdo, de critério e de construto.

Predomínio da validade de conteúdo – 1º período (1900-1950)

Nessa época, estavam em voga as teorias da personalidade e com elas predominava o interesse pelos traços de personalidade (tipos, temperamentos, traços, aptidões etc.). Essas teorias (Psicanálise, Fenomenologia, Gestaltismo etc.) apresentavam em geral pouca fundamentação empírica, assumindo um caráter bastante nebuloso, quando não fantasioso. Nessa atmosfera, os testes dos traços eram considerados válidos na medida em que seu conteúdo correspondesse ao conteúdo dos traços teoricamente definidos pela teoria psicológica em questão.

Fora alguns poucos (teste de Binet-Simon, de Raven, de Thurstone e alguns testes projetivos ainda em voga), as dezenas de testes criados nessa época já fazem parte de uma história passada e podem ser ditos representantes da pré-história dos testes psicológicos.

Predomínio da validade de critério – 2º Período (1950-1970)

Prevalecia em Psicologia o enfoque do Behaviorismo Skinneriano, que influenciou também a Psicometria. Os testes eram concebidos como uma amostra de comportamentos e tinham como função predizer outros comportamentos ou comportamentos futuros. Um teste era, consequentemente, válido se predizia com precisão os comportamentos em uma futura ou em outra condição. Esse se tornava, assim, o critério de validade do teste. Não interessava saber por que o teste predizia algo, bastava mostrar que de fato ele o fazia e isso era o critério de sua validade. Esse modo de conceber os testes ainda persiste hoje em dia, mas parece que aos poucos sua

relevância vem se tornando secundária, tornando-se tão somente uma etapa, juntamente com a validade de conteúdo, no processo de elaboração dos testes psicológicos (ANASTASI, 1986).

Esse período se caracteriza por uma acentuada fuga do pensar teórico que definia a época anterior. O teste não era mais construído para representar traços de personalidade, e os itens (tarefas) eram selecionados a partir de um grande elenco (*pool of items*) que parecia se referir àquilo para o qual se queria uma medida, fazendo uso praticamente exclusivo e *a posteriori* de análises estatísticas, especialmente da correlação. Não era mais a teoria psicológica e sim a estatística que definia a qualidade do teste. Esse processo de empirismo cego se assemelha ao pescador que lança a rede não importa onde para ver o que pode colher e, em cima do colhido, decidir o que quer. No processo, tipicamente se perdem "toneladas" de itens puramente por não satisfazerem critérios estatísticos (KURTZ, 1948; CURETON, 1950; PRIMOFF, 1952). A atitude dos psicometristas dessa época é explicada por razões históricas, eles queriam se desfazer do que lhes parecia um teorizar gratuito e fantasioso do início do século XX em Psicologia. Contudo, já na década de 1970, os psicometristas procuravam voltar a um teorizar psicológico mais relevante e em cima dele elaborar seus testes, o que deu início ao terceiro período na concepção dos testes e de sua validade.

Predomínio da validade de construto – 3º Período (1970)

Esse período teve suas fontes históricas no artigo de Cronbach e Meehl (1955) sobre o modelo trinitário da validade (conteúdo, critério, construto). Eles próprios já diziam que a validade de construto exigia um novo tipo de teorizar em Psicometria. Entretanto, o impacto prático dessa visão dos autores só se faria sentir após os anos 1970. Na verdade, a volta à teoria psicológica em Psicometria se deve a vários fatores; entre eles, salientam-se os seguintes.

1) Preocupação com o desenvolvimento da teoria da personalidade e, em especial, da inteligência, com maior base empírica e valendo-se sobretudo das técnicas da análise fatorial (COMREY, 1970; GUILFORD, 1967; JACKSON, 1974; MILLON, 1983; CATTELL, 1965; CATTELL; STICE, 1957; CATTELL; WARBURTON, 1967).

2) Realização de estudos dos processos cognitivos (STERNBERG, 1977, 1984; STERNBERG; DETTERMAN, 1986; STERNBERG; RIFKIN, 1979).

3) Realização de estudos do processamento da informação (NEWELL; SHAW; SIMON, 1958; NEWELL; SIMON, 1958).

4) Insatisfação com os resultados do uso de testes na educação e no trabalho. Na clínica, ainda se utilizavam bastante os testes projetivos, em que predominava, aliás, o pensamento da primeira época dos testes, os quais se baseavam nas teorias dos traços de personalidade.

5) O impacto da Teoria de Resposta ao Item (TRI) por sua insistência no traço latente. A influência decisiva dessa teoria ocorre somente após os anos 1980, devido ao atraso na área da informática para fazer uso prático das análises estatísticas complexas que tal enfoque exige.

Na validação dos instrumentos psicológicos, a preocupação agora se concentra na validade de construto ou de traços latentes. Não está ainda finalizada a disputa entre a ênfase nos traços ou a ênfase nas situações (*construct-centered versus task-centered*) ou, como diz Messick (1994), entre a avaliação *task-driven versus construct-driven*. Parece, entretanto, que o conceito de validade dos testes psicológicos irá finalmente se reduzir à validade de construto, e o conteúdo e o critério serão apenas aspectos desta (ANASTASI, 1986; MESSICK, 1989, 1994; EMBRETSON, 1983; WIGGINS, 1989; CRONBACH, 1989; o qual, já em 1955, de algum modo, previa tal desenlace). Essa tendência é obviamente favorecida também pelos psicólogos da linha cognitivista (STERNBERG, 1985, 1990; GARDNER, 1983).

Retoma-se, agora, a validade dos testes. Nos manuais de Psicometria, costuma-se dizer que um teste é válido quando de fato mede o que supostamente deve medir. Embora essa definição pareça uma tautologia, na verdade ela não é, uma vez considerada a teoria psicométrica sobre o traço latente, exposta neste trabalho. O que se quer dizer com a definição é que, ao se medirem os comportamentos (itens), que são a representação do traço latente, está-se medindo o próprio traço latente. Tal suposição é justificada se a representação comportamental for legítima. Essa legitimação somente é possível se existir uma teoria prévia do traço que fundamente que a tal representação comportamental constitui uma hipótese dedutível da teoria. A validade do teste (este constituindo a hipótese) será, então,

estabelecida pela testagem empírica da verificação da hipótese. Pelo menos, esta é a metodologia científica. Assim, fica muito estranha a prática corrente na Psicometria de se agrupar intuitivamente uma série de itens e, *a posteriori*, verificar estatisticamente o que eles estão medindo. A ênfase na formulação da teoria sobre os traços foi muito fraca no passado. Com a influência da Psicologia Cognitiva, essa ênfase felizmente está voltando ou deverá voltar ao seu devido lugar na Psicometria.

Aliás, a Psicometria Clássica entende "aquilo que supostamente deve medir" como sendo o "critério", este representado por teste paralelo. Assim, "aquilo que" é o traço latente na concepção cognitivista da Psicometria e é o critério (escore no teste paralelo) na visão comportamentalista.

Diz Anastasi (1986, p. 3) que o processo de validação de um teste "inicia com a formulação de definições detalhadas do traço ou construto, derivadas da teoria psicológica, pesquisa anterior, ou observação sistemática e análises do domínio relevante do comportamento. Os itens do teste são então preparados para se adequarem às definições do construto. Análises empíricas dos itens seguem, selecionando-se finalmente os itens mais eficazes (i.e., válidos) da amostra inicial de itens".[1]

A validação da representação comportamental do traço, isto é, do teste, embora constitua o ponto nevrálgico da Psicometria, apresenta dificuldades importantes em três níveis ou momentos do processo de elaboração do instrumento; a saber, níveis da teoria, da coleta empírica da informação e da própria análise estatística da informação.

No nível da teoria, se concentram talvez as maiores dificuldades. Na verdade, a teoria psicológica se encontra ainda em estado embrionário, destituída quase que totalmente de qualquer nível de axiomatização, o que resulta em uma pletora de teorias, muitas vezes até contraditórias. Basta lembrar de teorias como Behaviorismo, Psicanálise, Psicologia Existencialista, Psicologia Dialética e outras que existiram simultaneamente e postularam princípios irredutíveis entre as várias teorias e pouco concatenados dentro de uma mesma teoria ou, então, em número insuficiente para se poder deduzir hipóteses úteis para o conhecimento psicológico. Com essa confusão no

1 A questão da elaboração de testes psicológicos é detalhadamente tratada em Pasquali, 2010.

campo teórico dos construtos, torna-se extremamente difícil para o psicometrista operacionalizá-los, isto é, formular hipóteses claras e precisas para testar ou, então, formular hipóteses psicologicamente úteis. Ainda quando a operacionalização for um sucesso, a coleta de informação empírica não será isenta de dificuldades, como, por exemplo, a definição inequívoca de critérios com os quais estes construtos possam ser idealmente estudados. Mesmo em nível das análises estatísticas, são encontrados problemas. Pela lógica da elaboração do instrumento, a verificação da hipótese da legitimidade da representação dos construtos se faz por análise do tipo fatorial (confirmatória), por meio da qual se procura identificar, nos dados empíricos, os construtos previamente operacionalizados no instrumento. Mas acontece que a análise fatorial faz algumas postulações fortes que nem sempre se coadunam com a realidade dos fatos. Por exemplo, essa análise assume que as respostas dos sujeitos aos itens do instrumento são determinadas por uma relação linear destes com os traços latentes. Há, ainda, o grave problema da rotação dos eixos, a qual permite a demonstração de um número sem fim de fatores para o mesmo instrumento.

Entretanto, infelizmente, a história da validade dos testes psicológicos é ainda uma área pelo menos obscura. Duas definições já demostram isso.

1) Validade pode ser mais bem definida como a extensão para a qual certas inferências podem ser feitas com base em escores de um teste ou em outras medidas (MEHRENS; LEHMANN, 1984).

2) Validade consiste na extensão para a qual um teste é verídico, preciso ou relevante ao medir um traço que pretende medir.

A primeira definição se coaduna com o pensar de Cronbach e Mehel (1955), da rede nomológica, depois sistematizada por Samuel Messick (1989) sob validade de construto. A segunda definição melhor se coadunaria com uma visão dualista do ser humano, na qual traço significa processos mentais. De qualquer forma, as duas definições se concentram no que veio a ser denominado de construto. Agora, o que é um construto?

Na visão de Messick e da grande maioria dos psicometristas atuais, o construto é uma ficção; ou melhor, são racionalizações, na expressão de Messick, pois para ele a validade de um teste consiste num julgamento

integrado e avaliativo do grau em que evidências empíricas suportam a adequação e a propriedade de inferências e ações fundamentadas nos escores de testes ou outras formas de avaliação. Na visão da Psicologia Cognitiva e das neurociências, construto é uma realidade mental.

As duas visões apresentam problemas gigantescos. Primeiramente, construto como ficção implica irracionalidade epistemológica: pois o construto é a causa dos comportamentos; mas estes são reais, então como é possível se conceber uma ficção (construto) que possa produzir uma realidade? Por outro lado, construto como realidade mental implica o conhecimento desta (sua estrutura e funcionamento); mas a Psicologia não conhece nem a estrutura nem o funcionamento de tais processos mentais. Essa concepção de construto salva a racionalidade epistemológica, mas nos coloca no reino do incógnito. Contudo, ela permite e justifica a procura e a pesquisa desses processos, tarefa que as neurociências vêm tentando realizar. Fica a sensação de que as duas versões nos deixam, no presente, "em um mato sem cachorro".

Diante de tamanhas dificuldades, os psicometristas recorrem a uma série de técnicas para viabilizar a demonstração da validade dos seus instrumentos. Fundamentalmente, essas técnicas foram as seguintes.

1) Na visão clássica – Técnicas que podem ser reduzidas a três grandes classes (modelo trinitário): as que visam a validade de construto, as que visam a validade de conteúdo e as que visam a validade de critério (APA, 1954).

2) Na visão atual – Técnicas que objetivam procurar evidências de validade com base em:

   - conteúdo;
   - processos de resposta;
   - estrutura interna;
   - relação com outras variáveis;
   - consequências da testagem (AERA; APA; NCME, 1999).

A visão atual é a predominante em Psicologia, embora ela tenha finalmente perdido totalmente o conceito de construto (COLLIVER; CONLEE; VERHULST,

2012), focalizando a validação dos testes psicológicos por meio do acúmulo de provas circunstanciais (ditas evidências de validade) para legitimar as decisões tomadas com base nos escores (às vezes chamada de validade consequencial porque foca nas consequências que se tiram a partir dos escores dos testes e não mais no construto). Mesmo assim, essas associações de Psicologia deixaram de fora um tipo de validade que sobretudo as neurociências vêm insistindo ser importante, a validade ecológica. A seguir será apresentado um pouco de tudo isso, tomando a opinião de Aera, Apa e NCME (1999) simplesmente como organograma.

Validade com base no conteúdo

Refere-se ao conceito tradicional de validade de conteúdo. Um teste tem validade de conteúdo se ele constitui uma amostra representativa de um universo finito de comportamentos (domínio). É aplicável quando se pode delimitar *a priori* e com clareza um universo de comportamentos, como é o caso dos testes de desempenho, que pretendem cobrir um conteúdo delimitado por um curso programático específico.

Para viabilizar um teste com validade de conteúdo, é preciso que se façam as especificações dele antes da construção dos itens. Essas especificações comportam a definição de três grandes temas: *i*) definição do conteúdo; *ii*) explicitação dos processos psicológicos (dos objetivos) a serem avaliados; e *iii*) determinação da proporção relativa de representação no teste de cada tópico do conteúdo.

Quanto ao conteúdo, trata-se de detalhá-lo em tópicos (unidades) e subtópicos e de explicitar a importância relativa de cada tópico dentro do teste. Tais procedimentos evitam as indevidas super-representação de alguns tópicos e sub-representação de outros por vieses e pendores pessoais do avaliador. Claro que será sempre o avaliador ou a equipe de avaliadores que vai definir esse conteúdo e a relativa importância de suas partes, mas essa definição deve ser estabelecida antes da construção dos itens, a fim de garantir certa objetividade pelo menos nas decisões.

Quanto aos objetivos, um teste não deve ser elaborado para avaliar exclusivamente um processo. Como na aprendizagem entram em ação vários

processos psicológicos, há interesse em todos, ou naqueles que se quer que sejam avaliados por um teste de conteúdo. Por exemplo, o teste deverá conter itens que avaliam a memória (reproduzir), a compreensão (conceituar, definir), a capacidade de comparação (relacionar) e a capacidade de aplicação dos princípios aprendidos (solução de problemas, transferência da aprendizagem).

A validade de conteúdo de um teste é praticamente garantida pela técnica de construção deste. Assim, é importante esboçar essa técnica. Ela comporta os seguintes passos.

1) Definição do domínio cognitivo
   Definir os objetivos ou os processos psicológicos que se quer avaliar. Para essa tarefa, é útil se inspirar em alguma taxonomia clássica de objetivos educacionais, como, por exemplo, a taxonomia de Bloom (1956). Com base em uma taxonomia, definem-se os objetivos gerais e específicos que se deseja medir no teste, como:

   - conhecer tais tópicos;
   - compreender tais tópicos;
   - aplicar tais tópicos;
   - analisar tais tópicos.

2) Definição do universo de conteúdo
   Como o teste constitui uma amostra representativa do conteúdo, é preciso definir e delimitar o universo do conteúdo programático em divisões e subdivisões (tópicos e subtópicos) ou quantas outras subclassificações forem necessárias. Isso implica delimitar o conteúdo em suas unidades e subunidades de ensino.

3) Definição da representatividade de conteúdo
   Definir a proporção com que cada tópico e subtópico deve ser representado no teste, decidindo, assim, a importância com que cada um deles aparece no conteúdo total do universo.

4) Elaboração da tabela de especificação
   Nela são relacionados os conteúdos com os processos cognitivos a serem avaliados, bem como a importância relativa a ser dada a cada unidade.

5) Construção do teste
Elaborar os itens que irão representar o teste seguindo as técnicas de construção de itens (MAGER, 1981; PASQUALI, 2010).

6) Análise teórica dos itens
Essa análise visa verificar a compreensão das tarefas propostas no teste por parte dos testandos (análise semântica) e a avaliação da pertinência do item à unidade correspondente, bem como o processo cognitivo envolvido (análise de juízes).

7) Análise empírica dos itens
Após a aplicação do teste, os dados obtidos podem ser utilizados para validação empírica deste, para seu uso futuro. Essa análise implica basicamente na determinação dos níveis de dificuldade e de discriminação dos itens. A técnica da teoria da resposta ao item (TRI) pode ser de grande valia nessa etapa.

Para facilitar a especificação do teste, pode-se utilizar uma tabela de dupla entrada, com o detalhamento dos objetivos (processos) à esquerda, o detalhamento dos tópicos no topo, e, no corpo da tabela, o número de itens.

Validade com base nos processos de resposta

> Alguns estudos mais recentes fazem uma análise teórica-empírica das relações entre os processos mentais ligados ao construto em causa e as respostas aos itens do instrumento. A partir de propostas explicativas dos processos mentais subjacentes às respostas aos itens, formulam-se modelos explicativos sobre como a pessoa processa as informações dos itens do teste. A partir disso tenta-se prever aspectos da resposta como acertos e tempo de reação a diferentes itens em razão das suas características e demandas consequentes aos processos cognitivos ou emocionais. Assim busca-se analisar a coerência entre as explicações teóricas e os dados empíricos (NUNES; PRIMI, 2010, p. 122).

Contudo, é difícil ver nisso demonstração de validade do teste; trata-se de uma relevante curiosidade de estudo da Psicologia Cognitiva, mas não de uma prova de validade.

Validade com base na estrutura interna

Entre as cinco fontes de validade dos testes nos padrões de Aera, Apa e NCME (1999), este tipo de validade seria o único que poderia salvar o conceito de construto. Numa visão cognitivista de construto, a validade de construto ou de conceito é considerada a forma mais fundamental de validade dos instrumentos psicológicos e com toda a razão, dado que ela constitui a maneira direta de verificar a hipótese da legitimidade da representação comportamental dos traços latentes e, portanto, se coaduna exatamente com a teoria psicométrica aqui defendida. Historicamente, o termo construto entrou na Psicometria por meio da American Psychological Association Committee on Psychological Tests, que trabalhou entre 1950 e 1954 e cujos resultados se tornaram as recomendações técnicas para os testes psicológicos (APA, 1954).

O conceito de validade de construto foi elaborado com o clássico artigo de Cronbach e Meehl (1955) *Construct validity in psychological tests*, embora o conceito já tivesse uma história sob outros nomes, tais como validade intrínseca, validade fatorial e até validade aparente (*face validity*). Essas várias terminologias demonstram a confusa noção que construto possuía. Embora tenham tentado clarear o conceito de validade de construto, Cronbach e Meehl ainda o definem como a característica de um teste enquanto mensuração de um atributo ou de uma qualidade, o qual não tenha sido " definido operacionalmente". Reconhecem, entretanto, que a validade de construto reclamava por um novo enfoque científico. De fato, definir essa validade do modo que eles a definiram parece um pouco estranho em ciência, dado que conceitos não definidos operacionalmente não são suscetíveis de conhecimento científico. Conceitos ou construtos são cientificamente pesquisáveis somente se forem, pelo menos, passíveis de representação comportamental adequada. Do contrário, serão conceitos metafísicos e não científicos. O problema é que, sintetizando a atitude geral dos psicometristas da época, para definir validade de construto, os autores partiram do teste, isto é, da representação comportamental, em vez de partir da teoria psicológica que se fundamenta na elaboração da teoria do construto (dos traços latentes). O problema não é descobrir o construto a partir de uma representação existente (teste), mas sim descobrir se a representação (teste) é legítima, adequada do construto. Esse enfoque exige uma colaboração, bem mais estreita do que existe, entre os psicometristas

e a Psicologia Cognitiva. A validade de construto de um teste pode ser trabalhada sob vários ângulos: a análise da representação comportamental do construto, a análise por hipótese, a curva de informação da TRI, além do falsete estatístico do erro de estimação da Teoria Clássica dos Testes.

Erro de estimação

Essa forma de avaliar a validade de um teste era típica da Psicometria Clássica. Esse é um modelo de psicometria que poderia ser chamado de positivista, uma vez que ele se fundamenta exclusivamente nos dados empíricos coletados de um conjunto de itens agrupados inicialmente mais ou menos de maneira intuitiva. Na verdade, o teste (conjunto de itens) é construído mediante seleção de uma amostra de itens coletados de um universo de itens que parecem medir um dado construto. Essa maneira de construir instrumentos psicométricos se fundamenta na ideia de que existe, para cada construto, um universo indefinido de itens (*pool of itens*), do qual uma amostra é extraída para constituir o teste. Como é que se sabe inicialmente que os itens incluídos na amostra se referem a um construto somente ou que estamos retirando itens de um universo unidimensional para compor o teste? Apela-se aqui à famosa ou malfadada validade aparente (*face validity*), isto é, os itens parecem estar se referindo à mesma coisa! Por mais estranho que isto pareça ser, honestamente, é o que se faz na tradição positivista da Psicometria. É que nessa tradição falta todo o teorizar prévio sobre o construto (traço latente) para o qual se quer construir o instrumento de medida. Sem os procedimentos teóricos sobre o traço latente, os itens não são construídos para representá-lo comportamentalmente, mas são coletados mais ou menos a esmo ("chutados"), com base na validade aparente, e verificados depois, por meio de análises estatísticas, para ver se de fato eles estão ou não se referindo a alguma coisa (construto) comum. Assim, a Psicometria se torna, no máximo, um ramo da Estatística, como, aliás, era normalmente definida, e não um ramo da Psicologia, como deve ser concebida. Para a Estatística, número é número, não interessa de onde ele vem; mas para a Psicologia (Psicometria) o número é uma representação de conteúdo psicológico, então interessa muito de onde ele vem. Na tradição clássica da Psicometria, apela-se demasiadamente à Estatística para salvar a teoria psicológica. Isso não se aplica. Não se pode abdicar da teoria psicológica em favor da Estatística. É preciso, primeiramente, desenvolver e avançar a teoria psicológica (dos traços latentes) e apelar, em seguida, à Estatística para auxiliar na tomada

de decisões mais objetivas sobre a demonstração de hipóteses psicologicamente significativas e relevantes, estas deduzidas da teoria psicológica e não levantadas intuitiva e aleatoriamente. A Psicometria Clássica, e também a moderna, necessita urgentemente da ajuda da Psicologia Cognitiva neste particular, a fim de que possa instrumentalizar-se com a teoria dos traços latentes, para os quais ela quer desenvolver instrumentos de observação quantitativa (medida).

De qualquer forma, também na TCT se procura demonstrar a validade dos testes. Como é que isso era feito?

Nesse contexto, a Psicometria Clássica procura legitimar a validade de um instrumento segundo o conceito de erro de estimação, isto é, quanto o escore obtido pelo sujeito no teste se afasta do escore verdadeiro.

A fórmula para o cálculo do erro de estimação (EE), na qual um critério é predito com base em um teste, é a seguinte:

$$EE = s_C \sqrt{1 - r_{TC}^2}$$

na qual, s_c é o desvio-padrão da medida do critério e r_{TC}^2 é o coeficiente de validade, isto é, a correlação entre o teste e o critério.

Essa fórmula está fundamentada na ideia de se computar o erro mínimo que se pode cometer ao se predizer o escore de um teste a partir do escore de um teste paralelo.

Para poder obter o erro de estimação, é necessário possuir a medida de um critério, este que supostamente é a medida da aptidão. Por mais precário que tal procedimento pareça ser, é um dos poucos de que dispõe a Psicometria Clássica para estabelecer o erro de estimação e, por consequência, a validade de um teste, entendida como a precisão com a qual o teste pode predizer o escore verdadeiro. A fórmula deixa claro que, se o coeficiente de validade r_{TC}^2 for zero, então o erro de estimação será igual ao desvio-padrão da medida, pois o fator sob a raiz equivaleria a 1. Tal ocorrência implicaria que o teste não é capaz de predizer o escore verdadeiro melhor do que uma simples adivinhação, isto é, ele é totalmente inútil para predizer qualquer coisa. Agora, se o coeficiente de validade for

diferente de 0, então o teste tem poder maior de predizer do que uma simples adivinhação. Quanto maior? Se o coeficiente de validade fosse igual a 1, o erro de estimação seria 0, pois ele seria o desvio-padrão multiplicado por 0. Suponha-se que um teste tenha coeficiente de validade de 0,80, o que constitui um coeficiente de grandeza extraordinária em termos práticos, nesse caso, qual seria a força de predição do teste com respeito ao critério que pretende medir? Calculando o erro de estimação do teste, tem-se

$$EE = 1 \times \sqrt{1 - 0{,}80^2} = \sqrt{1 - 0{,}64} = \sqrt{0{,}36} = 0{,}60.$$

Assim, a predição do teste é 40% (1,00 – 0,60) superior à predição feita ao acaso ou por adivinhação. Isso não parece grande coisa dado um coeficiente de validade tão elevado, mas sempre é melhor que a pura adivinhação. Também, felizmente, o erro de estimação e o coeficiente de validade não são os únicos nem os melhores procedimentos para estabelecer a validade de um teste, como se verá a seguir.

Na verdade, há nesse procedimento do erro de estimação um certo equívoco ao se supor que o escore verdadeiro seja a medida daquilo que o teste pretende medir. De fato, o escore verdadeiro constitui um agregado de medida daquilo que o teste pretende medir mais as características peculiares dos itens que compõem o teste, sem que estas tenham a ver com o que este pretende medir.

Análise da representação

São utilizadas duas técnicas para demonstrar a adequação da representação do construto pelo teste: a análise da consistência interna e a análise fatorial.

Análise da consistência interna do teste

A análise da consistência interna consiste em calcular a correlação que existe entre cada item do teste e o restante dos itens ou o total (escore total) dos itens. Dado que o item analisado contribui para o escore total, ele teoricamente não deve entrar nesse escore, já que é ele que está sendo escrutinado. Assim, a correlação legítima será a do item com o restante dos itens. Essa preocupação é importante quando o número de itens do teste for pequeno, pois nesse caso o próprio item em análise afeta substancialmente o escore total a seu favor. Por exemplo, em um teste com 10 itens, cada

um contribui e influencia o escore total em 10%. Quanto maior, contudo, o número de itens que compõem o teste, menos relevante a influência de cada um em particular no escore total. Em um teste com 100 itens, por exemplo, cada um afeta o escore total em apenas 1%. Consequentemente, no caso de um teste com grande número de itens (n ≥ 30), a correlação do item com o escore total ou com o restante dos itens não vai fazer diferença relevante.

A análise da consistência interna do teste implica o cálculo das correlações de cada item individualmente com o restante do teste. Essa análise apresenta um problema lógico que se situa no escore total. Na verdade, o escore total é o critério contra o qual cada item é avaliado; mas acontece que os itens são os que vão constituir o escore total, antes mesmo de se saber se eles são válidos e somáveis (unidimensionais, isto é, que medem um e o mesmo traço latente). O escore total constitui, assim, uma dificuldade, dado que ele somente faz sentido se o teste já é *a priori* homogêneo. A correlação de cada item com o escore total já pressupõe que os itens são somáveis, isto é, homogêneos e válidos; em outras palavras, se pressupõe que todos os itens constituam uma representação adequada do traço e de um mesmo traço latente (unidimensionalidade). Além disso, a consistência interna pressupõe que os itens estejam intercorrelacionados, isto é, que as correlações entre eles mesmos sejam elevadas. Entretanto, as intercorrelações entre os itens não são uma demonstração de que estes estejam medindo o mesmo construto. Suponha a situação de três itens saturados em três fatores, como apresentados a seguir.

Tabela 1 Itens saturados em fatores

Item	F1	F2	F3
1	0,80	0,30	0,30
2	0,30	0,80	0,30
3	0,30	0,30	0,80

As correlações entre os três itens são todas de 0,57, altas e significativas, mas nem por isso se pode dizer que eles estejam medindo uma e a mesma coisa. Na verdade, o item 1 mede especificamente o fator 1, pois está altamente saturado somente neste fator e não nos outros dois, e os outros itens medem outros fatores. Consequentemente, a análise da consistência

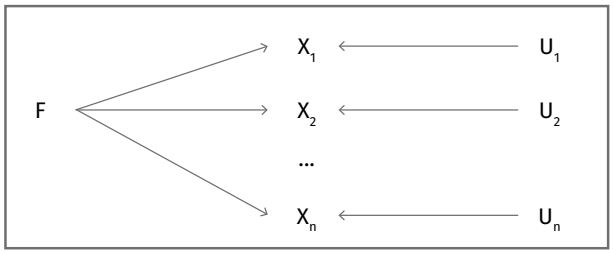
interna dos itens não parece garantir que eles sejam uma representação unidimensional de um construto.

A conclusão que se impõe dessas observações é a de que a análise da consistência interna não constitui prova cabal de validade de construto do teste.

Análise fatorial

Por outro lado, a análise fatorial tem como lógica precisamente verificar quantos construtos comuns são necessários para explicar as covariâncias (intercorrelações) dos itens. As correlações entre os itens são explicadas, pela análise fatorial, como resultantes de variáveis-fonte que seriam as causas dessas covariâncias. As variáveis-fonte são os construtos ou traços latentes de que fala a Psicometria. A análise fatorial também postula que um número menor de traços latentes (variáveis-fonte) é suficiente para explicar um número maior de variáveis observadas (itens), como se verifica na figura 1.

Figura 1 Representação do modelo fatorial



O modelo da figura 1 mostra que *n* variáveis (X) podem ser explicadas por um fator comum a todas as variáveis (F) e um fator específico para cada uma delas (U), de sorte que cada variável tem sua equação expressa em termos destes dois fatores.

Por exemplo:
$X_1 = a_1F + d_1U_1$.

O a1 é a saturação, a correlação, a covariância (dita carga fatorial) da variável X1 no fator F. Ela representa o percentual de relação que tem com o fator, isto é, quanto porcento ela se constitui em representação do fator

(traço latente); indica, em outras palavras, se ela é uma boa representação comportamental do traço latente. Além disso, as cargas fatoriais são as que determinam a correlação entre as próprias variáveis empíricas. Assim, a correlação entre X_1 e X_2 é definida por a_1a_2.

Dessa forma, a validade de construto de um teste é determinada pela grandeza das cargas fatoriais (que são correlações que vão de -1 a +1) das variáveis no fator, sendo aquelas a representação comportamental do fator, que, por sua vez, é o traço latente para o qual elas foram inicialmente elaboradas como representação empírica. Essas cargas fatoriais representam a parte fundamental do escore verdadeiro (V) da equação da Psicometria Clássica: T = V + E. Diz-se parte fundamental porque outra parte do V é constituída pela contribuição específica do item (contida no fator U do modelo fatorial) para o escore empírico T do teste. De fato, a variância total de um item ou variável pode ser decomposta em variância comum, variância específica e variância erro.

A variância comum representa o que as variáveis do teste têm em comum (expressa pelas intercorrelações entre elas) e que é recolhida nas cargas fatoriais no fator comum F. É esta que constitui a questão da validade do teste, isto é, quanto do traço latente (fator F) é representado empiricamente pelas variáveis (itens). O restante da variância dos itens é recolhido na chamada unicidade (U) de cada item que representa tanto o que é específico de cada um deles quanto os erros de medida. Estes dois últimos aspectos da variância (especificidade e erro) são agrupados em um conceito só, a saber, a unicidade, porque eles não contribuem para a validade do teste, pois é a porção do item que não constitui representação do traço latente.

Se não houvesse dificuldades com o modelo da análise fatorial, ele constituiria uma demonstração empírica cabal da validade de construto de um teste, pois forneceria a expressão exata de quanto o teste estaria representando o traço latente. Mas, infelizmente, a análise fatorial apresenta alguns problemas importantes. Duas razões são a preocupação principal neste particular. Primeiramente, o modelo fatorial se fundamenta em equações exclusivamente lineares entre variáveis e fatores. Embora seja rotineiro em Matemática tentar, em primeira aproximação, um modelo linear, parece difícil admitir que as intercorrelações empíricas entre os itens e a relação destes com os fatores (variáveis-fonte) possam ser todas reduzidas a

equações lineares. Isso é tanto mais plausível quando se observa que em quiçá nenhum campo da Psicologia e das ciências psicossociais em geral se encontram tais equações. Encontram-se, sim, equações logarítmicas, exponenciais e outras, isto é, equações não-lineares, como, por exemplo, nas leis da Psicofísica (leis de potência) e da análise experimental do comportamento (lei da igualação). Em segundo lugar, existe o grave problema da rotação dos eixos, para a qual não existe nenhum critério objetivo, a não ser a interpretabilidade psicológica (semântica) dos fatores. Essa ocorrência permite, em tese, a descoberta de qualquer fator que se queira, o que torna a solução extremamente arbitrária. Contudo, se o teste for construído via teoria psicológica de traços latentes e não a esmo (como a coleta de uma amostra de itens com base em um universo arbitrário deles, como é de praxe na construção de testes), tem-se um critério objetivo de rotação dos eixos em função dos traços latentes para os quais os itens foram inicialmente construídos como representação comportamental. Nesse caso, a análise fatorial será utilizada como teste de hipótese e não como pesca de hipóteses, assumindo, assim, a Estatística, como é legítimo, o papel de testagem de hipóteses psicológicas formuladas pela teoria psicológica e não o papel de criar ela (Estatística) as hipóteses psicológicas (*a posteriori*).

Análise por hipótese

Essa análise se fundamenta no poder de um teste psicológico ser capaz de discriminar ou predizer um critério externo a ele mesmo; por exemplo, discriminar grupos-critério que difiram especificamente no traço que o teste mede. Esse critério é procurado de várias formas. Há quatro entre as mais salientes e normalmente utilizadas, a saber, a validação convergente-discriminante, a idade, outros testes do mesmo construto e a experimentação.

A técnica da validação convergente-discriminante (CAMPBELL; FISKE, 1959) parte do princípio de que para demonstrar a validade de construto de um teste é preciso determinar duas coisas: *i*) o teste deve correlacionar significativamente com outras variáveis, com as quais o construto medido pelo teste deveria, pela teoria, estar relacionado (validade convergente); e *ii*) não se correlacionar com variáveis com as quais ele teoricamente deveria diferir (validade discriminante).

A idade é utilizada como critério para a validação de construto de um teste quando este mede traços que são intrinsecamente dependentes de

mudanças no desenvolvimento cognitivo/afetivo dos indivíduos, como é o caso, por exemplo, na teoria piagetiana do desenvolvimento dos processos cognitivos e da teoria de Spearman sobre a inteligência. A hipótese a ser testada nesse método é a de que o teste que mede o traço X, o qual muda claramente com a idade, é capaz de discriminar distintamente grupos de idades diferentes.

A prova que se faz nesse caso é a da diferença entre a média no teste de sujeitos mais jovens ($\overline{T}_j$) e a média de sujeitos mais adultos ($\overline{T}_a$), a saber

$$t = \frac{\overline{T}_a - \overline{T}_j}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_j^2}{n_j}}}$$

na qual, $\overline{T}_j$ e $\overline{T}_a$ são as médias no teste do grupo jovem e do grupo adulto, s_j^2 e s_a^2 são as variâncias destas médias e n_j e n_a são os números de sujeitos nos dois respectivos grupos.

Os graus de liberdade para verificar a significância do teste de Student "t" são $n_j + n_a - 2$.

Na história dos testes psicológicos, esse procedimento de validação foi talvez o primeiro a ser utilizado quando Binet e Simon (1905) empregaram o critério de diferenciação por idade na seleção dos itens do seu famoso teste de inteligência. Embora a preocupação explícita dos autores fosse construir um teste que fosse capaz de predizer o desempenho acadêmico de alunos do primeiro grau, eles se basearam numa hipótese de caráter conceitual, isto é, de que as habilidades cognitivas aumentam sistematicamente com a idade cronológica (na infância) e, para medi-las, escolheram tarefas específicas, cuja execução correta correspondia a determinada faixa etária.

Esse método contém um problema, o qual consiste no fato de que a maturação psicológica pode assumir dimensões e conotações muito distintas em culturas diferentes, por um lado; por outro, outras variáveis que não o traço em questão podem ser dependentes dessa maturação, dificultando ou impossibilitando a definição dos grupos-critério somente em função

da idade. Assim, se outras variáveis se alteram com a idade, pode bem ser que estas sejam as responsáveis pelas mudanças no escore e não a idade especificamente. Isso não seria um grave problema se essas outras variáveis covariassem sistematicamente com o traço latente que o teste quer medir e, além disso, variassem do mesmo modo em qualquer contexto cultural ou sócio-econômico, o que obviamente é difícil de assumir. Dentro de uma mesma cultura, o método pode se apresentar como importante para a determinação da validade de construto.

A correlação com outros testes que meçam o mesmo traço é também utilizada como demonstração da validade de construto. O argumento é de que, se um teste X mede validamente o traço Z e o novo teste N se correlaciona altamente com o teste X, então o novo teste mede o mesmo traço medido por aquele teste.

Essa técnica também contém um problema, o qual consiste no fato de que normalmente um teste de um traço qualquer não se apresenta com tal pureza a se poder afirmar que ele mede exclusivamente o tal traço. De fato, ele mede o traço em termos de um certo nível de covariância: por exemplo, existe uma correlação de 0,70 entre o teste X e o traço, o que equivale a uma comunalidade de 49% entre os dois. Agora, o novo teste N correlaciona 0,78 com aquele teste X. Há, portanto, comunalidade de 61% entre os dois testes. Qual será, nesse caso, a comunalidade do novo teste com o traço em si? Por azar poderia acontecer que a comunalidade de 61% entre os dois testes ocorra precisamente com os 51% do primeiro teste que não covariam com o traço; nesse caso, a comunalidade do novo teste com o traço seria de apenas 10%, isto é, o novo teste seria uma representação quase totalmente equivocada do traço.

O uso da intervenção experimental aparece logicamente como uma das melhores técnicas para se decidir a validade de construto de um teste. Essa técnica consiste em verificar se o teste discrimina claramente grupos-critério "produzidos" experimentalmente em termos do traço objeto de medida do teste. Assim, um teste que mede ansiedade teria validade de construto (ansiedade) se discriminasse grupo não-ansioso de grupo ansioso, definidos estes grupos em termos de manipulações experimentais: o ansioso, por exemplo, criado assim por meio de experiências provocadoras

de ansiedade. Na medida em que se puder garantir que as manipulações feitas nos grupos-critério atingirem exclusivamente o traço em questão, a testagem da hipótese é válida. Como, normalmente, essas manipulações supostamente de uma variável de fato podem afetar uma série de outras variáveis, sobretudo se as variáveis interagirem, fica confusa a decisão sobre em que especificamente os grupos-critério diferem e, consequentemente, fica inconclusiva a decisão sobre a hipótese de que o teste discrimina os grupos-critério exclusivamente em termos do traço que ele pretende medir. Podendo-se garantir que não ocorre tal alastramento das manipulações, a hipótese fica corretamente colocada.

Em conclusão, a técnica da validação de construto via hipótese, que, de um ponto vista da metodologia científica, se apresenta como a mais direta e óbvia, esbarra na dificuldade que existe na definição inequívoca do critério a ser utilizado como representante da manifestação do traço.

Deve-se, na verdade, concluir que todas estas técnicas de validação apresentam dificuldades. Nem por isso se justifica o simples abandono delas. Primeiramente porque em ciência empírica nada existe de perfeito e isento de erro e, em segundo lugar, a consciência das dificuldades deve servir para melhorar e não abandonar as técnicas. Aliás, é recomendável o uso de mais de uma das técnicas analisadas para demonstrar a validade de construto do teste, dado que a convergência de resultados das várias técnicas constitui garantia para a validade do instrumento.

Validade com base na relação com outras variáveis

Esse tipo de validação dos testes praticamente se confunde com o conceito tradicional de validade de critério.

Concebe-se como validade de critério de um teste o grau de eficácia que ele tem em predizer um desempenho específico de um sujeito. O desempenho do sujeito torna-se, assim, o critério contra o qual a medida obtida pelo teste é avaliada. Evidentemente, o desempenho do sujeito deve ser medido/avaliado mediante técnicas que são independentes do próprio teste que se quer validar.

Costuma-se distinguir dois tipos de validade de critério: *i*) validade preditiva e *ii*) validade concorrente. A diferença fundamental entre os dois tipos é basicamente com relação ao tempo que ocorre entre a coleta da informação pelo teste a ser validado e a coleta da informação sobre o critério. Se essas coletas forem (mais ou menos) simultâneas, a validação será do tipo concorrente; caso os dados sobre o critério sejam coletados após a coleta da informação sobre o teste, fala-se em validade preditiva. O fato de a informação ser obtida simultaneamente ou posteriormente à do próprio teste não é um fator tecnicamente relevante à validade do teste. Relevante, sim, é a determinação de um critério válido. Aqui se situa precisamente a natureza central desse tipo de validação dos testes, a saber: *i.* definir um critério adequado e *ii.* medir, de forma válida e independentemente do próprio teste, esse critério.

Quanto à adequação dos critérios, pode-se afirmar que há uma série deles que são normalmente utilizados, encontram-se listados a seguir.

1) Desempenho acadêmico – Talvez seja ou tenha sido o critério mais utilizado na validação de testes de inteligência. Consiste na obtenção do nível de desempenho escolar dos alunos, seja por meio das notas dadas pelos professores, seja pela média acadêmica geral do aluno, seja pelas honrarias acadêmicas que o aluno recebeu ou seja, até mesmo, pela avaliação puramente subjetiva dos alunos por parte dos professores ou colegas. Embora seja amplamente utilizado, esse critério tem igualmente sido muito criticado, não em si mesmo mas pela deficiência que ocorre na sua avaliação. É sobejamente sabida a tendenciosidade por parte dos professores em atribuir as notas aos alunos, tendenciosidade nem sempre consciente, mas decorrente de suas atitudes e simpatias em relação a este ou aquele aluno. Essa dificuldade poderia ser sanada até com certa facilidade se os professores tivessem o costume de aplicar testes de rendimento que possuíssem validade de conteúdo, por exemplo. Como essa tarefa é dispendiosa, o professor tipicamente não se dá ao trabalho de validar (validade de conteúdo) suas provas acadêmicas.

   Nesse contexto, é também utilizado como critério de desempenho acadêmico o nível escolar do sujeito: sujeitos mais avançados,

repetentes e evadidos. A suposição é de que quem continua regularmente ou está avançado academicamente em relação a sua idade possui mais habilidade. Evidentemente, nessa história não entra somente a questão da habilidade, mas muitos outros fatores sociais, de personalidade etc., o que torna o critério bastante ambíguo e espúrio.

2) Desempenho em treinamento especializado – Trata-se do desempenho obtido em cursos de treinamento em situações específicas, como no caso de atividades ligadas à música, à pilotagem, atividades mecânicas ou eletrônicas especializadas etc. No final desse treinamento, há tipicamente uma avaliação, a qual produz dados úteis para servirem de critério de desempenho do aluno. As observações críticas feitas ao ponto 1 valem também neste parágrafo.

3) Desempenho profissional – Trata-se, nesse caso, de comparar os resultados do teste com o sucesso/fracasso ou o nível de qualidade do sucesso dos sujeitos na própria situação de trabalho. Assim, um teste de habilidade mecânica pode ser testado contra a qualidade de desempenho mecânico dos sujeitos na oficina de trabalho. Evidentemente continua a dificuldade de levantar adequadamente a qualidade do desempenho dos sujeitos em serviço.

4) Diagnóstico psiquiátrico – Muito utilizado para validar testes de personalidade/psiquiátricos. Os grupos-critério são aqui formados em termos da avaliação psiquiátrica que estabelece grupos clínicos: normais *versus* neuróticos, psicopatas *versus* depressivos etc. Novamente, a dificuldade continua sendo a adequação das avaliações psiquiátricas feitas pelos psiquiatras.

5) Diagnóstico subjetivo – Avaliações feitas por colegas e amigos podem servir de base para estabelecer grupos-critério. É utilizada essa técnica sobretudo em testes de personalidade, nos quais é difícil encontrar avaliações mais objetivas. Assim, os sujeitos avaliam seus colegas em categorias ou dão escores em traços de personalidade (agressividade, cooperação etc.), com base na convivência que eles têm com os colegas. Nem precisa mencionar as dificuldades enormes que tais avaliações apresentam em termos

de objetividade; contudo, a utilização de um grande número de juízes poderá diminuir os vieses subjetivos nessas avaliações.

6) Outros testes disponíveis – Os resultados obtidos por meio de outro teste válido, que prediga o mesmo desempenho que o teste a ser validado, servem de critério para determinar a validade do novo teste. Aqui fica a pergunta óbvia: para que criar outro teste se já existe um que mede validamente o que se quer medir? A resposta se baseia numa questão de economia, isto é, utilizar um teste que demanda muito tempo para ser respondido ou apurado como critério para validar um teste que gaste menos tempo.

No caso desse tipo de validade, é preciso atender a duas situações bastante distintas. Primeiramente, quando existem testes comprovadamente validados para a medida de algum traço, eles certamente constituem um critério contra o qual se pode com segurança validar um novo teste. Infelizmente essa situação ocorre quase exclusivamente no caso da medida da inteligência, em que dispomos de alguns testes cuja validade já tem sido comprovada repetidas vezes, como é o caso das escalas de Wechsler (1975), de Stanford-Binet (TERMAN; MERRILL, 1960) e quiçá os dois fatores de inteligência fluida e cristalizada de Cattell (1971) e o fator G de Spearman (1927). Nos outros campos, há muita confusão. Talvez em relação à personalidade já existam alguns instrumentos válidos, como, por exemplo, o Questionário de Personalidade de Eysenck (*Eysenck Personality Questionnaire* – EPQ em EYSENCK; EYSENCK, 1975), no qual ele se refere às variáveis extroversão e neuroticismo ou ansiedade. O que vale aqui é o princípio de que se houver um teste comprovadamente válido para a medida de algum traço latente, ele certamente pode servir de critério para a validação de um novo teste. Espera-se nesse caso que a correlação do novo teste seja elevada em pelo menos 0,75.

Entretanto, quando não existem testes aceitos como definitivamente validados para avaliar algum traço latente, a utilização dessa validação concorrente é extremamente precária. Essa situação infelizmente é a mais comum. De fato, existem testes para medir praticamente "não importa o quê", como atestam os Buro's Mental Measurement Yearbooks, que são publicados periodicamente com centenas e milhares de testes psicológicos existentes no mercado. Nesse caso, pode-se utilizar esses testes como critérios de validação, mas o risco é demasiadamente

grande, porque se está utilizando como critério testes cuja validade é pelo menos duvidosa.[2]

Pode-se concluir que a validade concorrente só faz sentido se existirem testes comprovadamente válidos que possam servir de critério contra o qual se quer validar um novo teste e que esse novo teste tenha algumas vantagens sobre o antigo (como, por exemplo, economia de tempo etc.).

Contudo, uma pergunta frustrante fica ao final desta exposição sobre validade de critério. Se o pesquisador empregou toda a sua habilidade para construir um teste sob as condições de maior controle possível, por que iria ele validar essa tarefa-teste contra medidas inferiores, representadas pela medida dos vários critérios aqui apresentados. "Justifica-se validar medidas supostamente superiores por medidas inferiores?" – pergunta Ebel (1961).

Com as críticas de Thurstone (1952) e sobretudo de Cronbach e Meehl (1955), a validade de critério deixou de ser a técnica panaceia de validação dos testes psicológicos em favor da validade de construto. Contudo, os critérios apresentados acima podem ser considerados bons e úteis para fins de validação de critério. A grande dificuldade em quase todos eles se situa na demonstração da adequação da medida deles; isto é, em geral, a medida deles é precária, e, por isso, deixa muita dúvida quanto ao processo de validação do teste. Entretanto, há exemplos conhecidos de testes validados mediante esse método.

Validade com base nas consequências da testagem

Embora as consequências ou o uso dos escores de um teste não pareçam ter a ver com a validade dele (GREEN, 1998; MEHRENS, 1997; COLLIVER; CONLEE; VERHULST, 2012; CIZEK; BOWEN; CHURCH, 2010), as interpretações e o uso que se fazem dos escores dos testes adquiriram grande importância e consenso entre os pesquisadores e usuários dos testes para uso legítimo destes (KANE, 2006; PERIE; MARION; GONG, 2009 NICHOLS; WILLIAMS, 2009). Trata-se mais de responsabilidade social dos testes do que prova de sua validade como

2 No Brasil existe uma saída pragmática para isso: se o teste está com avaliação favorável no Satepsi, então ele é um bom teste.

medida. Isso porque o uso dos escores de um teste para tomar decisões de intervenção precisa legitimar tal ato. Assim, o uso de escores inadequados para tomar tais decisões em uma dada situação torna a atitude do psicólogo até criminosa, o que no final das contas vai respingar sobre a qualidade do teste do qual se extraíram os escores que fundamentaram as decisões. Enfim, é uma visão pragmática dos testes psicológicos na medida em que eles são utilizados para o bem-estar do ser humano. Tal intento é legítimo e necessário. Mas será essa atitude diante dos testes psicológicos, de se tornar a preocupação central da avaliação psicológica, útil para desenvolver o conhecimento e a teoria psicológica, dado que esta se fundamenta em inferências com base nos escores para processos mentais (construtos)? Mehrens (1997, p. 17) afirma: "Pode-se investigar a validade da inferência de que um escore seja um indicador razoável do montante do construto que possui independentemente de qualquer uso específico do escore". Como consequência, não se pode utilizar análises dos efeitos do uso do teste como evidência de sua validade. Enfim, incorporar as consequências de uso dos escores de um teste na demonstração de validade do teste se apresenta ainda como uma diatribe do tipo quixotesco. Kane (2013) confessa que o usuário do teste pode confundir a invalidade do uso do escore do teste com a invalidade do significado do escore, e Mehrens (1997, p. 18), por outro lado, afirma que "se validade é tudo, então validade não é nada".[3] Enfim, Cizek, Bowen e Church (2010) afirmam que as consequências da testagem como fonte de evidência de validade simplesmente não existem na literatura profissional e na medida aplicada, bem como em trabalhos de política na área. Os autores concluem que esses achados implicam a necessidade de se buscar, pelo menos, refinamentos na teoria e práticas atuais de validação dos testes.

De qualquer forma, quais são, então, as precauções a serem tomadas nesse contexto para salvaguardar a validade de um teste considerando-se que as consequências de uso dos escores do teste impactam a validade dele?

Em primeiro lugar e sobretudo, embora um teste possa ser utilizado para várias situações ou atividades, nenhum teste é adequado para todas as situações e atividades do ser humano. Assim, o teste deveria ser elaborado para situações ou atividades específicas, do que resulta que, no final das contas,

3 Para perceber a confusão nessa área, veja Michael T. Kane (2013), que procura mostrar, em um emaranhado discurso, o enfoque fundamentado em argumentação, a validação de um teste mediante a validação das interpretações e o uso dos escores do teste.

se deveria elaborar um teste diferente para cada situação ou atividade. É isso razoável ou possível? Green (1998) e Reckase (1998) opinam que impor tal tarefa de coletar evidência para as consequências de uso dos escores do teste sobrecarrega o seu criador com uma tarefa impossível. De qualquer forma, fica como responsabilidade dele mostrar para que situações ou atividades o teste produz escores adequados para a tomada de decisões com base nele.

Validade ecológica

Finaliza-se esse texto com a validade ecológica. Esta realmente não constitui uma nova forma de coletar evidências de validade, mas sim a forma como tais evidências devem ser buscadas. Ou seja, validade ecológica significa que os métodos, os materiais e as situações de um estudo dessa natureza devem se aproximar ao máximo do mundo real que está sendo examinado (BREWER, 2000).

Um exemplo: testar alunos na sala de aula. Se eles são assim acostumados, então a validade ecológica é alta, porque o processo não irá afetar o comportamento deles. Se, ao contrário, os alunos forem testados individualmente em uma sala isolada, então a validade ecológica cai drasticamente, porque não é o ambiente em que eles estão acostumados a ser testados nem a forma como costumam ser testados.

Assim, afirmar que validade ecológica consiste em tornar a situação de pesquisa similar aos fenômenos do mundo real é correto, mas é somente um aspecto da questão. Mark A. Schmuckler (2001) apresenta três critérios ou dimensões para iniciar uma compreensão adequada do que é validade ecológica, quais sejam: natureza do ambiente, dos estímulos e da resposta.

1) Natureza do ambiente de pesquisa – Brunswik (1943) iniciou esse debate criticando a artificialidade e o isolamento das situações de pesquisa (laboratórios) com respeito à realidade de vida dos sujeitos, pois eles não são representativos dos padrões amplos da vida. Nesse sentido, Bronfenbrenner (1977, p. 516; 1979) deu uma definição clássica de validade ecológica: “validade ecológica se refere à extensão na qual o ambiente experienciado pelos sujeitos na investigação científica possui as propriedades

que são da experiência dos sujeitos do experimento". Assim a representatividade e a naturalidade do ambiente de pesquisa constitui elemento fundamental da validade ecológica, isto é, o comportamento do indivíduo deve ser aferido em um ambiente verdadeiro em que os atores se comportam costumeiramente.

2) Natureza dos estímulos – Como no caso do contexto, aqui também vale o princípio da representatividade e da naturalidade, ou seja, os estímulos ou questões devem consistir em ocorrências atuais e estáveis do mundo real (naturalidade) e que sejam relevantes ao sujeito com respeito ao objeto de interesse a ser investigado (representatividade, importância). Isto é, os estímulos (questões) não devem ser esdrúxulos e extravagantes.

3) Natureza da tarefa, do comportamento ou da resposta – Novamente, a tarefa e a resposta pedida ao sujeito deve fazer parte de sua vida, de seu dia a dia, e não ter acontecido uma vez em sua vida. Bronfenbrenner (1977, p. 513; 1979) se revolta contra o que acha ocorrer na pesquisa em Psicologia do Desenvolvimento ao afirmar que ela é "a ciência do comportamento estranho das crianças em situações estranhas com adultos estranhos durante um período curtíssimo de tempo".

A validade ecológica cria tensões entre os pesquisadores, porque uns acham que atender às demandas desta torna a pequisa menos precisa (deficiência no controle das variáveis em jogo), enquanto outros argumentam que a artificialidade da pesquisa imposta pelo controle das variáveis em jogo torna os resultados irrelevantes para as situações reais da vida. A solução desse problema provavelmente se encontra no equilíbrio entre as duas preocupações. Algo similar foi observado por Campbell e Stanley (1973) no que diz respeito à preocupação com a validade interna e à preocupação com a validade externa das pesquisas científicas: "pesquisa sem validade interna produz somente erros; pesquisa sem validade externa produz resultados inúteis".[4]

4 *"Magni passus, sed extra viam"*, diriam os romanos ("Grandes passos, mas fora do caminho!").

Referências

AMERICAN PSYCHOLOGICAL ASSOCIATION (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, D.C.: American Psychological Association, 1954.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION (AERA), AMERICAN PSYCHOLOGICAL ASSOCIATION (APA), NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION (NCME). *Standards for psychological and educational testing*. Washington, D.C.: American Psychological Association, 1999.

ANASTASI, A. Evolving concepts of test validation. *Annual Review of Psychology*, v. 37, p. 1-15, 1986.

BINET, A.; SIMON, TH. Le développement de l'intelligence chez les efants. *Année Psychologique*, v. 14, p. 1-94, 1905.

BLOOM, B. S. *Taxonomy of educational objectives*: The classification of educational goals. Handbook I. Cognitive domain, New York: McKay, 1956. p. 201-207.

BREWER, M. B. Research design and issues of validity. In: REIS, H. T. (Ed.); JUDD, C. M. (Ed.). (2000). *Handbook of research methods in social and personality psychology*. New York, U.S.: Cambridge University Press, XII, 2000. p. 3-16.

BRONFENBRENNER, U. *The Ecology of Human Development*: Experiments by Nature and Design. Cambridge, MA: Harvard University Press, 1979.

BRONFENBRENNER, U. Toward an experimental ecology of human development. *American Psychologist*, v. 32, p. 515–531, 1977.

BRUNSWIK, E. Organismic achievement and environmental probability. *Psychological Review*, v. 50, p. 255-272, 1943.

CAMPBELL, D. T.; FISKE, D. W. Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, v. 6, p. 81-105, 1959.

CAMPBELL, D. T.; STANLEY, J. *Experimental and quasi–experimental design for research*. Skokie, IL: Rand McNally, 1973.

CATTELL, R. B. *The scientific analysis of personality*. Baltimore, MD: Penguin Books, Inc., 1965.

CATTELL, R. B . *Abilities:* their structure, growth and action. New York: Houghton Mifflin, 1971.

CATTELL, R. B.; STICE, G. F. *The Sixteen Personality Factor Questinonnaire* ("The 16 P.F."). Champaign, IL: Institute for Personality and Ability Testing, 1957.

CATTELL, R. B.; WARBURTON, F. W. *Objective personality and motivation tests*. Urbana, IL: University of Illinois Press, 1967.

CIZEK, G. J.; BOWEN, D.; CHURCH, K. Sources of Validity Evidence for Educational and Psychological Tests: A Follow–Up Study. *Educational and Psychological Measurement*, v. 70, n. 5, p. 732-743, 2010.

COLLIVER, J. A.; CONLEE, M. J.; VERHULST, S. J. From test validity to construct validity... and back? *Medical Education in Review*, v. 46, p. 366-371, 2012.

COMREY. A. L. *The Comrey Personality Scales*. San Diego, CA: Educational and Industrial Testing Service, 1970.

CRONBACH, L. J. Construct validation after thirty years. In: LINN (Org.), *Intelligence: Measurement, theory and public policy* – Proceedings of a symposium in honor of Lloyd G. Hamphreys, Chicago, IL: University of Chicago Press, 1989.

CRONBACH, L. J.; MEEHL, P.E. Construct validity in psychological tests. *Psychological Bulletin*, v. 52, p. 281-302, 1955.

CURETON, E. E.. Validity, reliability and baloney. *Educational and psychological measurement*, v. 10, p. 94-96, 1950.

EBEL, R. L. Must all tests be valid? *American Psychologist*, v. 16, n. 10, p. 640-647, 1961.

EMBRETSON, S. E. Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, v. 93, p. 179-197, 1983.

EYSENCK, H. J.; EYSENCK, S. G. B. *The Eysenck Personality Questionnaire*. Sevenoaks: Hodder; Stoughton, 1975.

GARDNER, H. *Frames of Mind*. New York: Basic Book Inc., 1983.

GREEN, D. R. Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement*, v. 17, n. 2, 16-19, 1998.

GUILFORD, J. P. *The nature of human intelligence*, New York: McGraw–Hill, 1967.

JACKSON, D. N. *Personality Research Form*. New York: Research Psychologists Press, 1974.

KANE, M. T. Validation. In: BRENNAN, R. L. (Ed.), *Educational measurement*. 4th ed. Washington, D.C.: The National Council on Measurement in Education & the American Council on Education, 2006, p. 17-64.

KANE, M. T. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, v. 50, p. 1-73, 2013.

KURTZ, A. K. A research test of the Rorschach test. *Personal Psychology*, v. 1, p. 41-51, 1948.

MAGER, R. F. *Medindo os objetivos de ensino ou "conseguiu um par adequado"*. Porto Alegre: Editora Globo, 1981.

MEHRENS, W. A. The Consequences of Consequential Validity. *Educational Measurement:* Issues and Practice, v. 16, p. 16-18, jun. 1997. DOI: 10.1111/j.1745–3992.1997.tb00588.x.

MEHRENS, W. A.; LEHMANN, I. J. *Measurement and evalution in education and psychology*. 3rd ed. New York: Holt, Rinehart, & Winston, 1984.

MESSICK, S. V. In: LINN, R. L. (Ed.), *Educational measurement*. 3rd ed. New York: Macmillan, p. 13-103, 1989.

MESSICK, S. V. The interplay of evidence nad consequences in the validation of performance assessments. *Educational Researcher*, v. 23, n. 2, p. 13-23, 1994.

MILLON, T. *Millon clinical Multiaxial Inventory Manual*. 2nd ed. Minneapolis, MN: National Computer Systems, 1983.

NEWELL, A., SHAW, J. C.; SIMON, H.A. Elements of a theory of human problem solving. *Psychological Review*, v. 65, p. 151-166, 1958.

NEWELL, A.; SIMON, H. A. Simulation of cognitive processes: A report on the summer research training institute. *Items*, v. 12, p. 37-40, 1958.

NICHOLS, P. D.; WILLIAMS, N. Consequences of test score use as validity evidence: Roles and responsibilities. *Educational Measurement:* Issues and Practice, v. 28, p. 3-9, 2009.

NUNES, C. H. S. S.; PRIMI, R. Aspectos técnicos e conceituais da ficha de avaliação dos testes psicológicos. In: CONSELHO FEDERAL DE PSICOLOGIA (Org.). *Avaliação psicológica: diretrizes na regulamentação da profissão.* Brasília: CFP, 2010. p. 101–128.

PASQUALI, L. *Instrumentação Psicológica*. Brasília, DF: Editora Vetor, 2010.

PERIE, M.; MARION, S.; GONG, B. Moving Toward a Comprehensive Assessment System: A Framework for Considering Interim Assessments. *Educational Measurement*: Issues and Practice, v. 28, p. 5-13, 2009. DOI:10.1111/j.1745–3992.2009.00149.x.

PRIMOFF, E. S. Job analysis attests to rescue trade testing from make–believe and shrinkage. *American Psychologist*, v. 7, p. 386, 1952.

RECKASE, M. D. Consequential Validity From the Test Developer's Perspective. *Educational Measurement*: Issues and Practice, v. 17, p. 13–16, 1998. DOI:10.1111/j.1745–3992.1998.tb00827.x.

SCHMUCKLER, M. A. What is ecological validity? A dimensional analysis. *Infancy*, v. 2, p. 419-436, 2001. DOI:10.1207/S15327078IN0204_02.

SPEARMAN, C. *The abilities of a man*. New York: MacMillan, 1927.

STERNBERG, R. J. Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities. Hillsdale, NJ: Erlbaum, 1977.

STERNBERG, R. J. General intellectual ability. In STERNBERG, R. J. (ed.) *Human abilities*: An information–processing approach. New York: Freeman, 1984. p. 5-29.

STERNBERG, R. J.; DETTERMAN, D. K. *Human intelligence*: Perspectives on its theory and measurement. Norwood, NJ: Ablex, 1986.

STERNBERG, R. J.; RIFKIN, B. The development of analogical reasoning processes. *Journal of Experimental Child Psychology*, v. 27, p. 196-232, 1979.

STERNBERG, R. J. *Beyond IQ:* A Triarchic Theory of Intelligence. Cambridge: Cambridge University Press, 1985.

STERNBERG, R. J. *Metaphors of mind: Conceptions of the nature of intelligence*. New York: Cambridge University Press, 1990.

TERMAN, L. M.; MERRILL, M. A. *Stanford–Binet Intelligence Scale*: Manual for the third revision, Form L–M. Boston: Houghton Mifflin, 1960.

THURSTONE, L. L. The criterion problem in personality research. *Psychometric Lab. Rep.*, IL: University of Chicago, Chicago, n. 78, 1952.

WECHSLER, D. Intelligence defined and undefined: A realistic appraisal. *American Psychologist*, v. 30, p. 135-139, 1975.

WIGGINS, J. S. A true test: Toward more authentic and equitable assessment. *Phi Delta Kappa*, v. 79, p. 703-713, 1989.

Luiz Pasquali
Doutor em Psicologia pela Université Catholique de Louvain, Bélgica
Professor da Universidade de Brasília
luiz.pasquali@gmail.com

USING ITEM MAPPING TO EVALUATE ALIGNMENT BETWEEN CURRICULUM AND ASSESSMENT

USANDO O MAPEAMENTO DE ITENS PARA AVALIAR O ALINHAMENTO ENTRE O CURRÍCULO E A AVALIAÇÃO

USO DE LA ASIGNACIÓN DE ELEMENTOS PARA EVALUAR LA ALINEACIÓN ENTRE EL PLAN DE ESTUDIOS Y LA EVALUACIÓN

Leah T. Kaira

Stephen G. Sireci

ABSTRACT

In educational testing, it is critical that the content of a test is aligned with the curriculum the test is designed to measure. Most methods for evaluating test-curriculum alignment rely on the subjective judgment of content made by experts who focus on how well the items on a test match curricular objectives. However, it is also important to ensure educational test items align with their expected levels of difficulty, which is much harder for experts to judge. In this study, test-curriculum alignment was evaluated by assessing the degree to which observed item difficulty aligned with intended item difficulty as determined by the test specifications. Using student response data for the Massachusetts Adult Proficiency Test (MAPT) for math, Item Response Theory (IRT) was used to locate items on the proficiency scale using two criterion response probability (RP) values. Item mapping results were compared to the item writers' classifications of the items, and degree of agreement between the two sets of data were statistically compared. In general, higher alignment was observed using RP50 than RP67, and for items assessing lower cognitive levels. Subject matter experts concluded cognitive demand, item clarity, and language complexity were viable reasons for misalignment.

Keywords: alignment; item mapping; Item Response Theory; response probability; validity.

RESUMO

Na avaliação educacional, é importante que o conteúdo do teste esteja alinhado com o currículo que ele pretende avaliar. A maioria dos métodos para avaliar o alinhamento entre o teste e o currículo tem como base o julgamento subjetivo do conteúdo feito por especialistas, que avaliam o quanto os itens do teste correspondem aos objetivos propostos no currículo. Não obstante, também é relevante garantir que os itens estejam alinhados com o nível de dificuldade esperado para eles, o que é mais difícil para os especialistas julgarem. Nesse estudo, o alinhamento entre o teste e o currículo foi verificado por meio da avaliação do grau com que o nível de dificuldade observado está de acordo com o esperado, conforme as especificações do teste. Para tanto, foram utilizadas as respostas dos estudantes ao Teste de Proficiência em Matemática para Adultos de Massachusetts (MAPT). A Teoria de Resposta ao Item (TRI) foi utilizada para localizar os itens na escala de proficiência usando os valores de dois critérios de probabilidade de resposta (RP). Os resultados do mapeamento dos itens foram comparados com a classificação feita pelos elaboradores e o grau de concordância entre os dois conjuntos de dados foram comparados estatisticamente. Em geral, um maior alinhamento foi observado usando RP50 do que RP67, e para itens que avaliavam níveis cognitivos mais baixos. Os especialistas concluíram que a demanda cognitiva, a clareza do item e a complexidade da linguagem foram as razões mais prováveis para o desalinhamento.

Palavras-chave: alinhamento; mapeamento de itens; Teoria de Resposta ao Item; probabilidade de resposta; validade.

RESUMEN

En la evaluación educativa es importante que el contenido de la prueba esté alineado con el currículo que se pretende evaluar. La mayoría de los métodos para evaluar la conformidad entre la prueba y el currículo tiene como base el juicio subjetivo del contenido hecho por especialistas, que evalúan cuánto corresponden los puntos de la prueba con los objetivos propuestos en el currículo. No obstante, también es relevante garantizar que los puntos estén alineados con el nivel de dificultad esperado para ellos, lo que para los especialistas es más difícil juzgar. En este estudio, la correspondencia entre la prueba y el currículo fue verificado por medio de la evaluación del grado con que el nivel de dificultad observado está de acuerdo con lo esperado, según las especificaciones de la prueba. Para esto, fueron utilizadas las respuestas de los estudiantes al Test de Competencia en Matemática para Adultos de Massachusetts (TCMA). Se utilizó la teoría de respuesta al ítem

(TRI) para ubicar los ítems en la escala de competencia usando los valores de los criterios de probabilidad de respuesta (PR). Los resultados del mapeo de los ítems fueron comparados con la clasificación hecha por los elaboradores, y el grado de concordancia entre los dos conjuntos de datos, se compararon estadísticamente. En general, una mayor alineación se observó al usar RP50, del que RP67, y para puntos que evaluaban niveles cognitivos más bajos. Los especialistas concluyeron que la demanda cognitiva, la claridad del punto y la complejidad del lenguaje fueron las razones más probables para la falta de alineación.

Palabras clave: alineación; mapeo de ítems; Teoría de Respuesta al Ítem; probabilidad de respuesta; valores.

Introduction

In educational testing, accurate evaluation of student learning can be achieved only if there is agreement among the curriculum, what the students learn, and what appears on the assessment. Similarly, assessment results are useful for accountability purposes if the assessment mirrors the curriculum. One strategy for evaluating the match between a curriculum and the assessment designed to measure it is carrying out alignment studies. Bhola et al. (2003, p. 21) define alignment as "the degree of agreement between a state's content standards for a specific subject and the assessment(s) used to measure student achievement of these standards".

Alignment is closely related to the interpretations made from test scores. According to the Standards for Educational and Psychological Testing (the Standards) (AMERICAN EDUCATIONAL RESEARCH ASSOCIATION [AERA], AMERICAN PSYCHOLOGICAL ASSOCIATION [APA], & NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION [NCME], 2014, p. 11), validity is "the degree to which evidence and theory support the interpretations of test scores entailed by proposed use of tests". Validation is therefore a process of collecting evidence to support the type of inferences that are drawn from test scores. Results of an alignment study can thus be used as validity evidence to support the interpretation of test scores.

The goal of alignment is to establish the degree of match between test content and subject area content as specified in curriculum standards. It

is important to emphasize the expression 'degree of agreement' because, as La Marca et al. (2000, p. 18) noted, "It is improbable that a single assessment instrument will provide the breadth of coverage necessary for an aligned system".

Breadth and coverage of curriculum standards necessitates test items that vary in difficulty. Some alignment methods, such as the Achieve Method (ACHIEVE, 2001), go beyond test-curriculum alignment by also evaluating "challenge," that is the degree to which the observed difficulty levels of items match their expected difficulty levels. In evaluating the "level of challenge" of items on a test, reviewers determine whether the sets of assessment items span an appropriate range of difficulty for students in the target grade level.

Current alignment methods

There has been an increase in research aimed at developing methodology for assessing alignment. According to Bhola et al. (2003), alignment methods can be categorized as having a low, moderate or high complexity based on level of focus, that is, the number of dimensions considered. For instance, a low complexity alignment study would only focus on the match between content of the items and the standards, while a high complexity study would also consider other dimensions such as the match in depth of content and the match between the levels of emphasis placed on a particular content area in the curriculum and in the assessment. One implication of this categorization is that different alignment studies may come up with different results depending on the levels of focus employed. As such, results from alignment studies of the same assessment, but employing different levels of focus, cannot be meaningfully compared.

Almost all alignment methods use subject matter experts (SME) to ensure they clearly understand the standards, the alignment criteria, and the scales being used to judge alignment. While expert judgments are essential in various steps in educational assessment, it is well known that despite training, humans make errors of unknown magnitude in their judgment. For example, Bhola et al. (2003) noted that SMEs may be overly generous in the number of matches they envision. Apart from the financial resources and the time required to convene SMEs, having SMEs review each item and make judgments over multiple criteria can also be cognitively challenging.

Another problem with current alignment methods is a lack of consensus regarding what constitutes sufficient alignment. Ananda (2003b, p. 20) noted one reason for lack of consensus is "...when articulating expectations for what students should learn (what they should know and be able to do), it is common for states to have different levels of statements, ranging from more global statements ...to narrower more targeted statements clustered under the broader statement". Thus, choice of alignment method is partly dictated by the breadth of statements describing what students should learn. This outcome could pose problems in evaluating improvements in the assessment as measured by student achievement.

Some alignment methods, such as the Achieve (2001) method, try to evaluate the appropriateness of the range of difficulty of the items on an assessment and the grade level of the students the assessment is intended for. In this process, it is assumed that after some training the SMEs have a common understanding of the range of abilities of the students in the target grade, and that they can accurately judge the difficulty of the item for a target group. However, research has shown that it is difficult for SMEs to make accurate judgments about the difficulty of items (Impara & Plake, 1998; Plake et al., 2000; Ryan, 1968; Shepard, 1994; Plake & Impara, 2001).

A good example of this difficulty is the 1990 National Assessment of Educational Progress (NAEP) math standard setting study in which great variability was observed among SMEs in making item judgments, despite training. The United States General Accounting Office (1993) claimed that the instruction given to the SMEs during training was not sufficient to bring the SMEs to a common understanding of what students at different achievement levels should know and be able to do. As a result, each SME formulated their own definition of what a basic, proficient or advanced student can do, resulting in large variability of judgments among the SMEs. The consequence of this variability was cut scores that were largely disputed and viewed as not representative of the knowledge and skills of the students assessed.

With respect to current alignment studies, it seems that evaluating the alignment of test items to their intended difficulty levels is important, but that alignment studies that rely on SMEs' subjective judgments are not going to be effective. That is, a mismatch between the SMEs' understanding of the range of student abilities at the target grade, and what the students

can actually do, could lead to item difficulty alignment results that are erroneous and misleading.

Thus, it seems reasonable to consider other approaches for evaluating item difficulty alignment than methods that rely on subjective judgment. In particular, methods are needed that (a) account for student's actual performance on items, (b) reduce reliance on subjective human judgment, and (c) apply consistent criteria for evaluating alignment.

An item mapping approach to evaluate item difficulty alignment

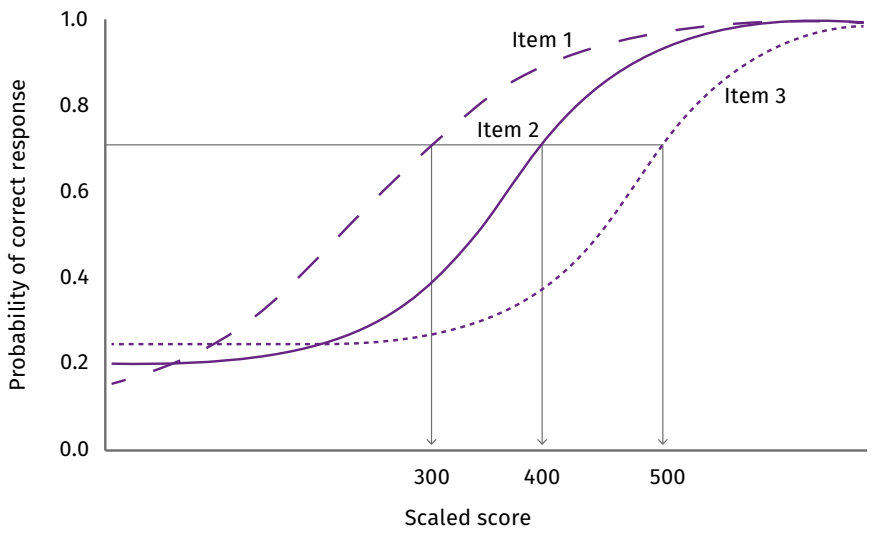
One method that could be used to evaluate the alignment of the intended and the observed difficulty of an item is item mapping. Item mapping has been widely used in educational assessment in areas of standard setting (e.g., Wang, 2003), scale anchoring (e.g., Gomez et al., 2006), and score reporting (e.g., Hambleton, 1997; Kirsch et al., 1993). Despite the various applications, the ultimate purpose of item mapping is to identify and describe what students at a specific level of achievement know and are able to do.

For the purposes of this study, item mapping is simply defined as the process of locating items along the test score scale. The idea behind item mapping is that given their characteristics, items could be systematically located on the test score scale based on some criteria. In most cases, the criterion used is the likelihood that examinees of a specified proficiency level have a high probability of success on the item.

The most popular approach for mapping items is the use of Item Response Theory (IRT). In IRT models, student achievement levels and item difficulties are on the same scale. Thus, given an examinee's proficiency, items the examinee would most likely answer correctly can be identified. The phrase 'most likely answer correctly' is usually defined by the probability that the examinee gives a correct answer to an item. This probability is referred to as the response probability (RP) criterion. In IRT models, each item is represented by an item characteristic curve (ICC), which gives the probability of correctly answering an item for a given proficiency level. Figure 1 shows ICCs for three dichotomously scored items. Item 3 has the lowest probability an examinee would give a correct response throughout most of the score scale. This implies that item 3 is more difficult compared to items 1 and 2.

Using a response probability of 70% (i.e., RP70), items 1, 2 and 3 would be mapped to scale scores of 300, 400, and 500 respectively. This means for example, that students with a scale score of 300 could be expected to correctly answer item 1 about 70% of the time. Similarly, students with scaled scores of 400 and 500 would be expected to correctly answer items 2 and 3, respectively, about 70% of the time.

Figure 1 Item characteristic curves for 3 hypothetical items



Application of item mapping to alignment

Item mapping could be used in an alignment study by locating items at specific points on the test score scale to help describe what students at that proficiency level can do. In cases where curriculum standards span several grade levels, and a vertical scale exists across those grade levels, the degree to which the items written for curriculum at higher grade levels are more difficult than items at lower grade levels can be evaluated. A system of tests that are aligned with respect to item difficulty will have items located along the IRT scale at locations that are implied by the test specifications. If items are located higher or lower with respect to their difficulties, some form of misalignment is present, and the source of the misalignment should be investigated. In this study, we show how item mapping can be used to

identify items that are misaligned with respect to their difficulty, and we use SMEs to evaluate reasons why the items are misaligned.

Purpose of current study

The purposes of our study are to investigate the utility of item mapping for evaluating the alignment between intended item difficulty (in terms of the grade span in which items are located) and actual item difficulty, and to discover reasons why misalignment in item difficulty may occur. The specific questions are:

- Can item mapping be used to enhance the evaluation of curriculum-assessment alignment from the perspective of item difficulty?
- Do response probability values have an impact on item difficulty alignment?
- If misalignment is observed, what are the likely causes?

Method

Empirical data were used to illustrate use of item mapping in assessing alignment among curriculum and assessment. The analyses were applied to data from a large-scale assessment in adult education.

Description of test

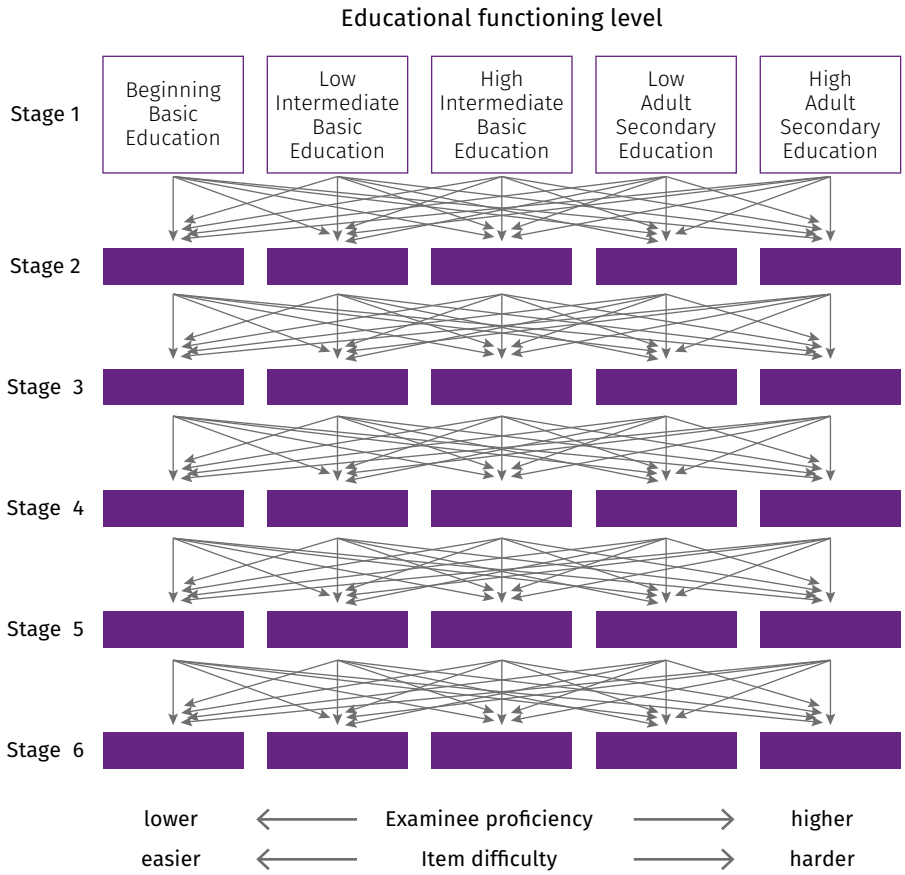
The Massachusetts Adult Proficiency Test (MAPT) for mathematics and numeracy is a computerized-adaptive multistage test (MST) designed to measure the mathematics achievement of adult education students in Massachusetts, USA. The MAPT for math is used by the State of Massachusetts to fulfill the Federal accountability requirements in adult education known as the National Reporting System (WASHINGTON, 2006). The NRS stipulates 6 achievement levels called educational functioning levels (EFLs), which are similar to grade levels in elementary through secondary school (i.e., each EFL spans about two grade levels). The MAPT measures all but the lowest EFL, and students taking the MAPT for the first time begin the test at an EFL designated by their teacher. Thus, there are 5 starting points for the MAPT. However, there are not separate tests for each EFL. Rather, all EFLs are calibrated along a common score scale and examinees may be routed to any EFL as they take the test, depending on how well they perform on the items administered at each stage.

The MST design for the MAPT involves six-stages, as illustrated in figure 2. The design includes two parallel panels, each consisting of 30 sets of items called modules. A panel is a collection of modules that defines all potential paths examinees may be routed to when taking the test (SIRECI et al., 2008). In MST, panels are analogous to alternate forms as defined in linear testing. The arrows in figure 2 show some (but not all) potential paths to which examinees may be routed. The first time a student takes the MAPT s/he is randomly assigned to one of the two panels. The other panel is used for a second test administration. A total of 40 scored items are administered to each student across the six stages. Students take 15 items during the first stage and 5 items in each of the subsequent stages. Proficiency estimates at each stage are used to determine the set of items (i.e., module) the examinee will take during the next stage. All items are dichotomously scored multiple-choice items with four answer choices.

The content of the MAPT for math is specified using two dimensions — one for test content and one for cognitive level. Four content areas are measured — Geometry and Measurement; Patterns, Functions and Algebra; Statistics and Probability; and Number Sense (hereafter referred to as Geometry; Patterns; Statistics; and Number Sense, respectively). The distribution of the items is 84, 68, 93, and 116 across the four content areas, respectively. With respect to cognitive level, three levels are specified — Knowledge and Comprehension; Application; and Analysis, Synthesis, and Evaluation. There were 114 items assessing Knowledge and Comprehension, 175 items assessing Application, and 73 items assessing Analysis, Synthesis and Evaluation. For convenience, the three cognitive skill areas will be referred to as Comprehension, Application, and Evaluation, respectively.

Each panel of the MAPT is designed to assess students' proficiency in math at five different EFLs: Beginning Basic, Low Intermediate, High Intermediate, Low Adult Secondary, and High Adult Secondary. There are separate test specifications for each EFL corresponding to the specific curricula for the EFL as described in the Massachusetts Adult Basic Education Curriculum Frameworks for Mathematics and Numeracy (SIRECI et al., 2008).

Figure 2 MST structure for the MAPT for math



Item mapping data

Response data for both panels for the 2009 administrations of the MAPT for math were used. About 7,361 examinees' responses to 362 math items were analyzed. The 3-parameter logistic IRT (3PL) model was used to estimate item parameters from examinee responses. Parameter estimation was done using BILOG-MG (ZIMOWSKI et al., 1996).

The items were also coded by the test developers with respect to content attributes including EFL, content strand (Geometry and Measurement; Patterns, Functions and Algebra; Statistics and Probability; and Number

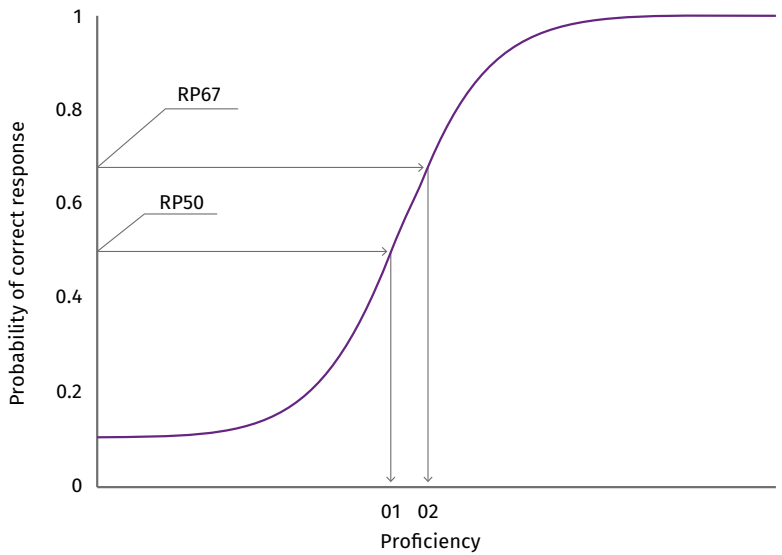
Sense), and cognitive skill (Knowledge and Comprehension; Application; and Analysis, Synthesis and Evaluation). The EFL designations for each item were originally made by the item writers, and subsequently confirmed by an independent group of SMEs. There were 100 Beginning Basic, 97 Low Intermediate, 94 High Intermediate, and 71 Low Adult Secondary items.

Item mapping method

We used a model-based item mapping method to identify items that mapped to a particular EFL. The steps were:

- Obtain parameter estimates for each item using the 3PL IRT model. We used the parameters for all items that were operational in 2009, with the exception of items in the High Adult Secondary EFL.
- Given the item parameter estimates, calculate the theta (θ) value required for an examinee to have some specified probability of correct response for each item. Figure 3 illustrates this estimation. The task is to find θ_1 and θ_2 for each item for which examinees have a probability of .50 (RP50) and .67 (RP67), respectively, of success on the item.
- Determine the EFL to which each item mapped. We used theta values obtained in step 2, and the cut scores for each EFL.

Figure 3 An item characteristic curve illustrating the Model Based Item Mapping Method



Response probability values

RP50 and RP67 were used to determine the EFL to which each item mapped and to assess the impact of RP value on the alignment results. RP50 and RP67 were chosen because these are the most common RP values in literature (KARANTONIS & SIRECI, 2006). Use of these RP values allowed for comparison of results of this study with findings of similar studies reported in literature. Second, because a goal was to illustrate how item mapping could be applied to evaluation of curriculum-assessment alignment, an operational definition of what students "can do" was needed. Based on literature, there seems to be a consensus that for tests that do not have very high stakes for individuals, RP values higher than 67 may be too high (U.S. GENERAL ACCOUNTING OFFICE, 1993), and certainly RP values lower than 50 cannot be used to claim students have mastered the concepts tested by an item.

An item was considered to map to a particular EFL if the probability of success on the item was .50 (for the RP50 condition) or .67 (for the RP67 condition) for examinees whose proficiency estimates (θ) were within the specified EFL. Each item was considered to map to the lowest level where examinees had a probability of providing a correct response at the RP or higher. After items were mapped to the various EFLs, results were compared to the test developer's classifications of the items. An item was considered to match or align to the intended EFL if the item mapping results agreed with the test developer's classification. A situation where an item is mapped to an EFL other than intended was considered a mismatch and misaligned.

Reasons for curriculum-assessment misalignment

Seven teachers were convened for a one-day meeting to look at 20 items that did not map as intended to find potential reasons to explain the misalignment. Stratified random sampling was used to selected the misaligned items to ensure that items from all EFLs are represented. The teachers came from all geographical locations across Massachusetts and were drawn from current ABE teachers. Seventy-one percent of the teachers were female and the rest were males. All teachers were Caucasian with teaching experience ranging from 3.5 to 32 years. The meeting began with self-introductions of the participants followed by training that the facilitator conducted. The training sessions began with communicating the goal of the meeting, which was to review items that mapped to higher or lower EFLs than the test

developers had intended and suggest reasons for the misalignment. The teachers were then given a set of 6 items, which were used as practice items. The teachers looked at the items, the objective and level it was intended for and tried to find reasons why the item did not map to the intended level (i.e., why it mapped to an easier or more difficult location than it was written to). The teachers first looked at the practice set of items individually, which was followed by a group discussion.

After training, the teachers were split into two groups: one group analyzed the items that were misaligned using the RP50 criterion, followed by items that were misaligned using the RP67 criterion; the other group followed the opposite order. The items were presented in two booklets with one booklet presenting items that misaligned at RP50 first and RP67 last, while the second booklet had the opposite ordering. Each teacher was presented with an item review sheet on which they recorded their reviews. Group discussions of some of the items followed individual review of the items. A questionnaire was administered to evaluate the item review process. This questionnaire contained 5 Likert-type and 2 open response items. The Likert-type questions were rated on a 5-point scale from strongly agree to strongly disagree. In general, the survey sought teachers' views on aspects of the meeting such as adequacy of time for item review, adequacy of training and clarity of the item review task. The open-ended questions asked the teachers about some factors that they used in coming up with possible reasons for the observed misalignment and suggestions for the future.

Data analyses

Results were analyzed to assess the degree of agreement between item mapping results and intended EFLs for each item. Comparisons across the intended and item-mapped classifications were made at the item, content strand, and cognitive skill levels. The comparisons involved examining the agreement between the item mapping results and the intended classifications for each RP value (R50 and RP67). Chi-square tests and correlations were used to assess the degree of alignment. Content analysis (Gall et al., 1996) was used to analyze written accounts provided by teachers.

Results

Overall item mapping results

A review of all the misaligned math items at RP50 revealed that in general, misaligned items were slightly more discriminating and harder than the aligned items. The average discrimination and difficulty parameter estimates were 1.49 and 0.68 respectively for misaligned items versus 1.33 and -0.23 respectively for the aligned items. The average pseudo-guessing parameter estimate was 0.2 for both groups of items. This observation may imply that both the a- and b-parameters had an impact on alignment results. The misaligned items selected for review had similar average discrimination parameter estimates to all misaligned items (1.51 vs. 1.49). However, the reviewed items were much harder ($\bar{b}$ =1.15 vs. 0.68) and had slightly lower average pseudo-guessing parameter estimates (0.17 vs. 0.20).

Table 1 presents the overall classifications of the math items based on RP50 and RP67. It can be seen that exact agreement (proportion of items mapping to intended level) for the beginning basic level was 34% using the RP50 criterion, and only 17% using the RP67 criterion. The "correct" mapping locations for the other EFLs tended to be lower. In all but one case – Low Adult Secondary (the highest level of items evaluated in the set) – the RP50 criterion located more items in the intended EFL than RP67. The majority of items were mapped to higher difficulty levels across all EFLs for both RP50 and RP67.

Table 1 Math overall item mapping results for RP50 and RP67

Intended MAPT level	% items mapped to level based on RP									
	BB		LI		HI		LAS		HAS	
	RP50	RP67	RP50	RP67	RP50	RP67	RP50	RP67	RP50	RP67
BB	**34.0**	**17.0**	37.0	33.0	24.0	31.0	5.0	17.0	0.0	2.0
LI	5.2	1.0	**28.9**	**9.3**	40.2	40.2	21.6	35.1	4.1	14.4
HI	0.0	0.0	11.7	3.2	**28.7**	**16.0**	33.0	28.7	26.6	52.1
LAS	0.0	0.0	2.8	0.0	16.7	2.8	**18.1**	**20.8**	62.5	76.4
Total	10.7	5.0	21.5	24.0	28.1	24.0	19.3	25.6	20.4	33.1

Notes: Shading indicates items mapped into intended levels. BB: Beginning Basic; LI: Low Intermediate; HI: High Intermediate; LAS: Low Adult Secondary; HAS: High Adult Secondary.

In looking at the RP50 results across all EFLs, the overall exact agreement between test developers' classification and IRT based item mapping at RP50 was 28.1%. This means only 28.1% of the items were mapped to the same level as intended. Combining exact and adjacent (items mapping to one EFL lower or higher than intended) agreement as a measure of overall agreement between test developers and item mapping classifications, overall agreement at RP50 was 77.5%. The highest adjacent agreement (84.8%) was obtained at the LAS level. The Spearman correlation between the RP50 classifications and intended classifications was 0.69, which is considered moderate based on Cohen's (1988) r2 criteria (r2 = .48). The chi-square test for these results was 234.66 (df = 15, p < .001) implying statistically significant differences exist between the PR50 item mapping results and the test developer's classifications of the items.

For RP67, the overall exact agreement between item mapping results and test developers classification was 15.4%, which is just over half the level of exact agreement for RP50. More items mapped to the LAS level or higher, relative to RP50. The highest exact agreement for RP67 was 20.8% at the Low Adult Secondary level. Only 36.6% of the items mapped to one EFL lower or higher based on the intended classification at RP67 compared to 47.1% for RP50. Overall agreement for RP67 was 59.5%. The highest adjacent agreement between the intended and IRT classification was 100% for the LAS level. The Spearman correlation between the RP67 classifications and the intended ELFs was 0.71 (r2 = 0.50), which was slightly higher than the correlation observed for RP50. Similarly, the chi-square results were statistically significant (χ^2_{15} =256. p < 0.001).

In summary, more congruence between the item mapping results and the classifications intended by the test developers was obtained at RP50. For both RP values, larger proportions of items map to one EFL higher than the EFL for which the item is intended. This may suggest that the items are generally harder than the test developers had anticipated.

Qualitative results: reasons for misalignment

Six broad categories pertaining to characteristics of items were derived from the reasons provided by the teachers during the study. The categories were: item difficulty, cognitive demand of the item, language level of the item compared to language level of the students, the type of math

concept being assessed, clarity of the item, and technical issues related to the item.

It was observed that the math concept being assessed in the item was a factor contributing to item difficulty misalignment in 13 items. The teachers noted that some mathematical concepts such as order of operations, calculating the mean in reverse order, finding the inverse, and math tasks involving mathematical symbols like greater than or less than, were generally harder for students. The teachers confirmed there were differences between the item developers' classifications and item mapping results due to some characteristics of the items that made them easier than intended. These characteristics included distractors that could be easily eliminated and familiarity of the scenario presented in the item.

The teachers identified cognitive demand of the item as a factor contributing to misalignment in 12 items. Most (9) items in this category asked students to derive and integrate new information into subsequent steps. The other items required students to extrapolate or perform multiple steps to arrive at the correct response.

With respect to why items may have been more difficult than expected, complexity of the language used in an item compared to reading level of the student was one factor that teachers suggested as contributing to misalignment in 11 items. Teachers noted that some items contained words that were hard for students at some EFLs and hence the poor performance on those items. For example, one teacher pointed out that reading and interpreting true/false statements was generally challenging for Beginning Basic students for whom English was a second language. Teachers noted that vocabulary such as doubling every minute, consistent, mean, inequality, average, perimeter, more than half, three times more, twice as often, and equivalent were hard for students to comprehend especially at the lower EFLs. The teachers also noted that some items contained long and complex sentences that required more sophisticated reading skills that students for whom the item was intended did not possess.

Eleven items were noted to exhibit some technical problems or ambiguities leading to students' poor performance. For example, in one item the stem

did not explicitly state that students needed to provide their answer in different units of measurement than the units in the stem. In another question, students were presented with a scenario where a fence needed to be put around a circular pond. However, the question did not specify that the fence also needed to be circular. Teachers cited lack of clarity of the item as a reason contributing to misalignment for 10 items. For instance, one teacher noted that in one item, students needed to reformulate the question to be able to answer it because the question was unclear. For one question, teachers noted the question was framed in such a way that it led examinees to carry out a wrong mathematical operation. Teachers also noted that presenting items in long sentences increased the likelihood of reducing the clarity of the item making the item become harder than intended. Similarly, teachers stated that some items contained information that was not necessary for students to respond to them and that may have led to confusion among some students.

A summary of the reasons teachers gave to explain why items were misaligned in terms of difficulty is presented in table 2. In addition to providing comments on individual items, teachers were also asked about the factors they considered in reviewing the item to generate possible reasons for misalignment. They cited language complexity, appropriateness of content for level of examinee, editorial errors in the question, and the number of steps required to solve the question. The teachers also mentioned the cognitive skill the item requires, the ability of examinees the item is intended to evaluate, and also the vocabulary used in the item as some of the factors they took into consideration.

Table 2 Summary of reasons for misalignment

Reason	Number of Items
Math concept assessed	13
Item difficulty	12
Cognitive demand	12
Language level	11
Technical issues with item	11
Item clarity	10

Discussion

This study was designed to illustrate how student responses to test items could be used to inform curriculum-assessment alignment. Item mapping based on IRT was applied to an adult basic education math assessment to illustrate the process. IRT was used to map the items in terms of their difficulty and the results were compared to test developers' classification of the items to evaluate the degree of agreement. SMEs (teachers) were used to help explain why some items were misaligned with respect to item difficulty.

The results of the present study indicate that some significant differences occurred between test developers' and item mapping classifications of the items. More items mapped to lower EFLs (that is High Intermediate or lower) at RP50 while more items mapped to higher EFLs (LAS or higher) at RP67. These results were expected because most of the items used in this study had c-parameter values that were less than 0.35. As such, the theta value at which students have a 50% chance of providing a correct response to an item (that is RP50) will always be less than the b-value. The only exception to this is when the c-parameter is equal to zero. On the other hand, the theta value at which students have a 67% chance of providing a correct response to an item will always be higher than the b-value. However, the assumption being made here is that the test developers took difficulty and discrimination of the item into consideration in classifying the items. The other assumption is that the test developers' estimation of the difficulty of the items for a particular group of learners was accurate. These assumptions are discussed later.

Results also showed that in general, greater alignment between test developers' and item mapping results was obtained at RP50 compared to RP67. These results are similar to results obtained by Kolstad et al. (1998). In their study aimed at evaluating the impact of RP value on selection of exemplar items for describing what students at a particular proficiency level could do, the authors found the greatest agreement between the percentage of items mapped along the proficiency scale and percentage of scores for examinees along the proficiency scale at RP50.

We used the degree of agreement between test developers' classifications of the items and item mapping results as a measure of a new kind of

alignment—alignment of item difficulty. This alignment evaluation is similar to the "level of challenge" criterion in the Achieve (2001) model, which uses subjective judgment. Alignment of item difficulty is important whenever a testing program uses a vertical scale across different grade levels. As the results of this study show, items do not always map to their intended levels. Such misalignment can only be discovered by analyzing students' responses to them.

Comparing agreement between test developers' classifications of the items and location of the items on the proficiency scale assumes that some common notion of difficulty was used in the two classifications. It is hoped that test developers consider not only the match between the item content and the level of the curriculum at which the content is taught, but also the relative difficulty of the item. As such, trustworthiness of test developer's ratings of the items for the intended group hinges upon their ability to accurately judge or estimate difficulty of the item for the target group.

Based on these results, it appears reasonable to state that test developers were not completely accurate in their estimates of the difficulty of the items. Results of this study closely match results obtained by Zwick et al. (2001), who investigated alternative item mapping methods for the NAEP. Zwick et.al asked SMEs to list the five easiest and the five hardest items from a test without ordering the items by difficulty within each set. They found that the SMEs difficulty rankings matched very closely to student's performance. Specifically, a Spearman correlation between the SMEs rankings and the proportion of 8th graders answering an item correctly was 0.65. Based on this correlation, Zwick et al. (2001, p. 22) concluded that the SMEs "rankings were substantially in line with the actual difficulty of the items" . Similar conclusions could be drawn about the math results obtained in the current study. Spearman correlations between test developers' and item mapping results were about 0.7 for both RP50 and RP67.

We also found that at RP50 greater exact agreement between test developers and item mapping results was obtained at the lower EFLs (i.e., High Intermediate level or lower) while the least was obtained at higher EFLs. Considering RP67, greater agreement between the two classifications was obtained at the higher EFLs compared to lower EFLs. These results imply that most items intended for lower EFLs mapped to low EFLs while those intended for higher EFLs did map to high EFLs. This finding also provides

some evidence that test developers made reasonably accurate judgments about items intended for lower EFLs and those intended for higher EFLs.

With respect to the SME teachers' reviews of the misaligned items, in general, items demanding higher levels of thinking were perceived to be more difficult. Analysis of the items that teachers identified as cognitively more demanding showed that they were those that the item writers classified as measuring evaluation and synthesis skills, which provides some validity evidence for the MAPT. Teachers identified difficult vocabulary, use of long sentences and excess verbiage as some factors contributing to misalignment. It was interesting to note that lack of student exposure to content was not a factor that contributed to low performance of examinees. It was the level of cognitive thinking the content in the item demanded that mattered most.

Test developers could improve alignment between intended and actual item difficulty by ensuring that language in the item matches language level of the students. This does not only improve the clarity of the item and student understanding but also eliminates construct irrelevant variance that could interfere with student performance. Another strategy would be to match cognitive demands of the item to cognitive capability of examinees. This would reduce the frustration and stress that might affect student performance in an item. Alignment can also be improved by ensuring that items are free from error. Items should be stated in simple language, and accompanying visuals should be well drawn and well labeled where appropriate. It is also important to ensure that the distractors are plausible, that is, they cannot be easily eliminated by less knowledgeable examinees or they do not offer clues to the correct response.

Implications of results

This study illustrated that the utility of alignment study results could be greatly enhanced if students' actual performance on the assessment can be taken into consideration. This would provide information on the strengths and weaknesses of the students and also inform teachers which areas of the curriculum need extra emphasis. Given that many large-scale assessment systems in K-12 education use assessments that are vertically equated across grades, the degree to which intended and actual item difficulty aligns is an important validity issue. The results of this study will help facilitate intended and actual alignment of item difficulty, and provide a method for helping evaluate it.

References

Achieve, Inc. *Measuring up* – a commissioned report on education assessments for Minnesota. Washington, D.C.: Author, 2001.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association, 2014.

Ananda, S. *Rethinking issues of alignment under No Child Left Behind*. San Francisco: WestEd, 2003a.

Ananda, S. Achieving alignment. *Leadership*, v. 33, n. 1. p. 18-21, 2003b.

Bhola, D. S. et al. Aligning tests with content standards: methods and issues. *Educational Measurement*: issues and practice, Washington, v. 22. p. 21-29, 2003.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. New Jersey: Lawrence Earlbaum, 1988.

Gall et al. *Educational Research:* An Introduction. 6th ed. New York: Longman Publishers, 1996.

Gomez, P. G. et al. Proficiency descriptors based on a scale-anchoring study of the new TOEFL iBT reading test. *Language Testing*, v. 24, n. 3, p. 417-444, 2007.

Hambleton, R. K. Enhancing the validity of NAEP achievement level score reporting. In: PROCEEDINGS OF ACHIEVEMENT LEVELS WORKSHOP. Washington, D.C.: National Governing Board, 1997.

Impara, J. C. & Plake, B. Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, v. 35, n. 1, p. 69-81, 1998.

Karantonis, A.; Sireci, S. G. The bookmark standard setting method: A literature review. *Educational Measurement*: issues and practice, v. 25, n. 1. p. 4-12, 2006.

Kirsch. I. et al. *Adult literacy in America*: A first look at the findings of the National Adult Literacy Survey. Washington, D.C.: National Center for Education Statistics; U.S. Department of Education, 1993.

Kolstad, A. et al. The response probability convention used in reporting data from IRT assessment scales: Should NCES adopt a standard? Washington, D.C.: American Institutes for Research, 1998.

La Marca, P. M. et al. *State Standards and State Assessment Systems*: A guide to alignment. Washington, D.C.: Council of Chief State Officers, 2000.

Plake, B. S. et al. Consistency of Angoff based predictions of item performance: Evidence of technical quality of results from the Angoff standards setting method. *Journal of Educational Measurement,* v. 37, n. 4. p. 347-355, 2000.

Plake, B. S.; Impara, J. C. *Ability of panelists to estimate item performance for a target group of candidates*: An issue in judgmental standard setting. Educational Assessment, v. 7, n. 2. p. 87-97, 2001.

Porter, A. C. Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, v. 31, n. 7. p. 3-14, 2002.

Ryan, J. J. Teacher judgment of test item properties. *Journal of Educational Measurement*, v. 5, n. 4. p. 301-306, 1968.

Shephard, L. A. Implications for standard setting of the NAE evaluation of NAEP achievement levels. In: JOINT CONFERENCE ON STANDARD SETING FOR LARGE SCALE ASSESSMENT. Washington, D.C.: National Assessment Governing Board, National Center for Educational Statistics, 1994.

Sireci, S. G. et al. *Massachusetts Adult Proficiency Tests technical manual*. Research Report n. 677, v. 2. Amherst, MA: University of Massachusetts, Center for Educational Assessment, 2008.

U.S. General Accounting Office. *Educational achievement standards*: NAGB's approach yields misleading interpretations. Report n. GAO-PEMD-93-12. Washington, D.C.: Author, 1993.

Wang, N. Use of the Rasch IRT model in standard setting: An item mapping method. *Journal of Educational Measurement*, v. 40 n. 3. p. 231-253, 2003.

Washington, D.C. U.S. Department of Education. Division of Adult Education and Literacy. Office of Vocational and Adult Education. *Implementation guidelines*: measures and methods for the national reporting system in adult education. Washington, D.C., 2006, July. Available at <http://www.nrsweb.org/foundations/implementation_guidelines.aspx>.

Zimowski, M. F. et al. BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items. Chicago: Scientific Software International, Inc., 1996.

Zwick, R. et al. An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement*: issues and practice, v. 20, n. 2. p. 15-25, 2001.

Leah T. Kaira
Doutora em Educação pela Universidade de Massachusetts Amherst, EUA
leahkaira@gmail.com

Stephen G. Sireci
Ph.D. em Psicometria pela Fordham University
Professor da Universidade de Massachusetts Amherst, EUA
sireci@acad.umass.edu

PROPOSTA DE SEGMENTAÇÃO DE UMA ESCALA DA TRI UTILIZANDO O NÍVEL SOCIOECONÔMICO[1,2,3]

PROPOSAL FOR SEGMENTATION OF A SCALE BY IRT USING THE SOCIOECONOMIC STATUS

PROPUESTA DE SEGMENTACIÓN DE UNA ESCALA DE LA TRI UTILIZANDO EL NIVEL SOCIOECONÓMICO

M.ª Gabriela Thamara de Freitas Barros

Adriano Ferreti Borgatto

Adolfo Samuel de Oliveira

RESUMO

Este trabalho apresenta uma proposta de segmentação de escala para um indicador de nível socioeconômico (Inse) construído com base no modelo politômico de respostas graduais da Teoria de Resposta ao Item (TRI) e nos questionários dos estudantes da Aneb, da Anresc e do Enem 2013. Na medida proposta, foram utilizadas apenas as informações pertinentes para segmentação da escala, que são os valores do Inse divididos em pequenas faixas e as probabilidades de resposta por item/categorias do modelo. Para definir o número de agrupamentos foi considerado o coeficiente de aglomeração e o gráfico de dendrograma. Esses procedimentos metodológicos facilitaram a formação de grupos (níveis), que foram descritos segundo os itens/as categorias de cada nível. Além disso, foi possível relacionar a medida criada com os itens que definem o indicador, facilitando a compreensão

1 Este trabalho é uma versão revisada e ampliada de um dos tópicos da dissertação de mestrado defendida no Programa de Pós-graduação em Métodos e Gestão em Avaliação da Universidade Federal de Santa Catarina, intitulada "Procedimentos para a construção de indicadores por meio da teoria de resposta ao item: a criação de uma medida de nível socioeconômico familiar" (BARROS, 2016).

2 Agradecemos ao Adriano Souza Senkevics e à Valéria Maria Borges, pesquisadores do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), pela consultoria dada na elaboração do *abstract* e do *resumen*, respectivamente.

3 A pesquisa de mestrado foi financiada pelo Inep, ao qual agradecemos pelo apoio.

do construto mensurado. Esses resultados evidenciaram as vantagens de construir uma escala empregando a TRI, bem como de segmentá-la.

Palavras chave: Teoria de Resposta ao Item; segmentação de escala; indicador de nível socioeconômico.

ABSTRACT

This work introduces a proposal of segmentation of a scale based on an Indicator of Socioeconomic Level (Inse), elaborated using the Item Response Theory (IRT) for polytomous response data and the student questionnaires of 2013 Aneb, Anresc and Enem evaluations, carried out by Inep. The proposed measure consisted of using only the information considered as important for the scale segmentation: the Inse values divided into small ranges and the response probabilities by item/category. Aiming to define the number of clusters, we took into consideration the clustering coefficient and the dendrogram plot. These methodological procedures were adopted to provide clues that could facilitate the formation of groups (levels), which, by their turn, were described from the items/categories of each level. In addition, we could associate the measure with the items that composed the indicator, making easier to understand the construct measured. These results pointed out the advantage of constructing a scale by using IRT, as well as segmenting it.

Keywords: Item Response Theory; segmentation of the scale; indicator of socioeconomic level.

RESUMEN

Este trabajo es una propuesta de segmentación de una escala basada en un indicador de nivel socioeconómico (Inse), que fue elaborado según el modelo politómico de respuestas graduales de la Teoría de Respuesta al Ítem (TRI) y los cuestionarios de los estudiantes de Aneb, Anresc y Enem 2013, aplicados por el Inep. La medida propuesta buscó utilizar sólo las informaciones pertinentes para la segmentación de escala, que son los valores del Inse divididos en pequeñas franjas y las probabilidades de respuesta por ítem/categorías del modelo. Con la finalidad de definir el número de agrupaciones, fueron considerados el coeficiente de aglomeración y el gráfico de dendrograma. Esos procedimientos metodológicos fueron escogidos para obtener direccionamientos que faciliten la formación de grupos (niveles), los cuales fueron descriptos según los ítems/categorías de cada nivel. Además, fue

posible relacionar la medida creada con los ítems que definen el indicador, lo que facilitó la comprensión del constructo valorado. Esos resultados evidenciaron la ventaja de construir una escala empleando la TRI, así como de segmentarla.

Palabras clave: Teoría de la Respuesta al Ítem; segmentación de escala; indicador de nivel socioeconómico.

Introdução

Os indicadores educacionais, especialmente a partir do atual Plano Nacional de Educação (BRASIL, 2014), tornaram-se um dos principais instrumentos de diagnóstico, monitoramento e avaliação do sistema de ensino brasileiro em seus diversos níveis de agregação, pois requerem a produção de informações sobre escolas e redes de ensino, considerando-se os três entes da federação.

No bojo desse movimento, a fim de contextualizar os resultados de suas avaliações e exames da educação básica, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), enquanto o maior responsável pela produção e divulgação de dados primários relacionados à educação escolar, construiu uma série de indicadores educacionais, dentre os quais se destacam os referentes à adequação da formação docente, às condições de trabalho e rotatividade dos professores, à complexidade da gestão escolar e ao nível socioeconômico das escolas (INEP, 2014a; 2014b; 2014c; 2014d; 2015). Acadêmicos do país também já vinham contribuindo com essa temática apresentando diversas propostas de indicadores sobre infraestrutura da escola e diversos outros fatores relacionados a ela (SOARES, 2005; BORTOLOTTI, 2012; NETO et al., 2013; ALVES; SOARES, 2013).

Muitos desses indicadores valeram-se da Teoria de Resposta ao Item (TRI) para construir suas escalas, com maior ou menor sucesso na comunicação de seus resultados em virtude da complexidade em traduzir um construto latente e abstrato em uma medida facilmente inteligível. De acordo com Jannuzzi (2001), a comunicabilidade é uma das propriedades desejáveis de um indicador, uma vez que o seu uso adequado e profícuo depende do nível de compreensão dos tomadores de decisão e dos implementadores de políticas e programas educacionais acerca do retrato que ele proporciona. Por esta razão,

mesmo que um indicador seja socialmente relevante, válido e fidedigno, se não for bem compreendido e utilizado corretamente pelos atores envolvidos na gestão educacional, ele acaba por não cumprir sua finalidade precípua.

Visando contribuir tanto com o aprimoramento da constituição da escala de um indicador de nível socioeconômico (Inse), construído com base no modelo politômico de respostas graduais da TRI, quanto com a sua comunicabilidade, é que se apresenta essa proposta de segmentação, pois, como destaca Franco (2004), a obtenção de medidas contextuais é um procedimento complexo e o campo das medidas sociais ainda tem o que se desenvolver no Brasil. Nesse sentido, essa tarefa é uma das mais importantes no campo educacional, pois, como ressalta o mesmo autor, não dá para formular políticas públicas sem considerar as medidas contextuais, razão pela qual, há, no País, uma série de medidas de acompanhamento do desenvolvimento socioeconômico que norteiam as políticas públicas, tais como o produto interno bruto (PIB), a renda *per capita*, o índice de desenvolvimento humano (IDH) e o Coeficiente de Gini.

No âmbito da educação, o nível socioeconômico familiar dos alunos já foi desenvolvido tanto por pesquisadores (SOARES; SOUZA; PEREIRA, 2004; SOARES, 2005; ALVES; SOARES, 2009; ALVES; SOARES; XAVIER, 2014; BARROS, 2016) quanto pelo Inep (2014d). De modo geral, esse construto é definido nesse trabalho como uma medida observada indiretamente e calculada operacionalmente pela agregação de informações sobre a educação, a ocupação e a riqueza ou rendimento dos alunos. Segundo Buchmann (2002), essa medida determina precisamente os resultados educacionais, motivo pelo qual investigar a forma como o aprendizado dos alunos se processa sem considerar o seu contexto socioeconômico e familiar pode levar a resultados imprecisos ou até equivocados.

Dentro desse quadro, a criação de indicadores é uma tarefa fundamental, embora seja difícil e muitas vezes exija a utilização de técnicas estatísticas avançadas, tal como a TRI, que é uma técnica bastante profícua, uma vez que possibilita medir o que não é diretamente observado e produzir uma escala interpretada e comparável entre as edições do indicador (ANDRADE; TAVARES; CUNHA, 2000). Nesse sentido, a TRI permite relacionar a medida calculada e os itens utilizados na criação do indicador, ou seja, possibilita interpretar qualitativamente a medida numérica criada.

No entanto, a descrição de cada valor do Inse, apesar de ser possível, não é viável, em virtude da quantidade de pontos de uma escala. Por essa razão, fazer cortes na escala numérica gerada (CIZEK, 2001; BEATON; ALLEN, 1992; HUYNH,1998), criando grupos característicos, de forma que ocorra uma descrição para cada grupo, com base em um valor de referência, tende a ser a melhor opção para solucionar esse problema (KLEIN, 2009). A esse respeito, duas questões são colocadas: *i)* como segmentar os indivíduos em grupos na escala; e *ii)* como interpretar esses grupos.

Alves, Soares e Xavier (2014) classificaram as escolas em sete grupos (mais baixo, baixo, médio baixo, médio, médio alto, alto e mais alto) de acordo com o nível socioeconômico médio de seus alunos. Para o agrupamento, utilizaram a técnica de análise de conglomerados – método de agrupamento não hierárquico *k-means*. Entretanto, os autores utilizaram essa técnica apenas para fazer a classificação das escolas em grupos, mas não fizeram a interpretação dos grupos em relação ao nível socioeconômico.

Em geral, observa-se que os pesquisadores optam por dividir os indivíduos em grupos, mas muitos não sugerem uma interpretação qualitativa da medida produzida. Tal situação se intensifica quando se trabalha com modelos politômicos, dada a dificuldade de trabalhar com uma maior quantidade de parâmetros para interpretar.

Tendo em vista esses desafios a respeito de como dividir e interpretar escalas produzidas com base em modelos politômicos da TRI, decidiu-se, neste trabalho, apresentar, utilizando o Inse calculado por Barros (2016), uma proposta alternativa ao que, no geral, foi encontrado na literatura até o momento, no que se refere a segmentação e interpretação de escala. Considera-se que essa proposta contribuirá para a busca de um ponto ótimo para segmentar a escala em relação às características dos itens.

Método

O Inse calculado por Barros (2016) foi construído utilizando as respostas dos questionários contextuais dadas pelos alunos participantes do Sistema de Avaliação da Educação Básica (Saeb) (composto pela Avaliação Nacional da Educação Básica – Aneb e pela Avaliação Nacional do Rendimento

Escolar – Anresc) e pelos concluintes do ensino médio que fizeram o Exame Nacional do Ensino Médio (Enem) em 2013. O estudo foi realizado utilizando inicialmente 24 itens a partir das respostas de 6.273.995 estudantes, sendo 2.524.125 do 5º ano do ensino fundamental do Saeb, 2.720.588 do 9º ano do ensino fundamental do Saeb, 150.429 da 3ª série do ensino médio do Saeb e 878.853 alunos do ensino médio do Enem. Procurou-se estudar um conjunto de modelos da TRI, de natureza politômica e dicotômica, a fim de utilizar aquele que melhor se ajustasse aos dados. Os testes mostraram que o modelo de respostas graduais (SAMEJIMA, 1969) foi o que melhor se ajustou aos dados, ensejando, assim, uma medida mais informativa para toda a escala, quando comparada com os demais modelos, e que possibilitou que os itens dos diferentes conjuntos de dados fossem calibrados em uma mesma escala.

A fim de explicitar o modelo de respostas graduais adotado, pode-se dizer que, no de Samejima (1969), as categorias do item são ordenadas, motivo pelo qual a probabilidade de um aluno *j* optar por uma categoria *k*, quando *k*= 0, 1, 2..., é dada por:

$$P^{+}_{i,k}(\theta_j) = \frac{1}{1+e^{-a_i(\theta_j - b_{i,k})}},$$

e a probabilidade de um aluno *j* escolher a categoria *k* no item i é dada por:

$$P_{i,k}(\theta_j) = P^{+}_{i,k}(\theta_j) - P^{+}_{i,k+1}(\theta_j).$$

Nesse modelo, portanto, o parâmetro *ai* representa a discriminação do item *i* e designa quanto cada item está relacionado com o construto nível socioeconômico; o parâmetro *bi,k* mensura o grau de dificuldade da categoria *k* do item *i* na escala do nível socioeconômico; e, por fim, os valores de *θj* são os valores do nível socioeconômico de cada aluno.

O Inse criado por Barros (2016) também verificou os pressupostos desse modelo, tais como a invariância e a unidimensionalidade, o que convergiu para um modelo final que utilizou dezessete itens (quadro 1). Além disso,

validou o Inse correlacionando-o com outras medidas similares, no que diz respeito a esse construto.

Quadro 1 – Itens e alternativas utilizadas na construção do indicador

Q02 – TV por assinatura em casa 1 – Não tem; 2 – Sim, uma; 3 – Sim, duas; 4 – Sim, três ou mais
Q03 – Videocassete e/ou DVD em casa 1 – Não tem; 2 – Sim, um; 3 – Sim, dois; 4 – Sim, três ou mais
Q05 – Internet em casa 1 – Não tem; 2 – Sim, uma; 3 – Sim, duas; 4 – Sim, três ou mais
Q06 – Rádio em casa 1 – Não tem; 2 – Sim, um; 3 – Sim, dois; 4 – Sim, três ou mais
Q08 – Telefone celular em casa 1 – Não tem; 2 – Sim, um; 3 – Sim, dois; 4 – Sim, três ou mais
Q09 – Aspirador em pó em casa 1 – Não tem; 2 – Sim, um ou mais
Q10 – Geladeira em casa 1 – Não tem; 2 – Sim, uma; 3 – Sim, duas; 4 – Sim, três ou mais
Q11 – Freezer (aparelho independente ou parte da geladeira) em casa 1 – Não tem; 2 – Sim, um ou mais
Q12 – Máquina de lavar roupa em casa 1 – Não tem; 2 – Sim, uma; 3 – Sim, duas ou mais
Q13 – Automóvel em casa 1 – Não tem; 2 – Sim, um; 3 – Sim, dois; 4 – Sim, três ou mais
Q14 – Banheiro em casa 1 – Não tem; 2 – Sim, um; 3 – Sim, dois; 4 – Sim, três ou mais
Q15 – Quartos para dormir em casa 1 – Não tem; 2 – Sim, um; 3 – Sim, dois; 4 – Sim, três; 5 – Sim, quatro ou mais
Q19 – Contrata empregado(a) doméstico(a) em casa 1 – Não; 2 – Sim, um ou mais

Q20 – Renda mensal da família (considerando a renda do aluno) 1 – Nenhuma renda; 2 – Até um SM; 3 – Mais de um até 1,5 SM; 4 – Mais de 1,5 até 3 SM; 5 – Mais de 3 até 7 SM; 6 – Mais de 7 até 9 SM; 7 – Mais de 9 até 12 SM; 8 – Acima de 12 SM
Q21– Mãe, ou a mulher responsável por você, sabe ler e escrever 1 – Não; 2 – Sim
Q22– Pai, ou homem responsável por você, sabe ler e escrever 1 – Não; 2 – Sim
Q25 – Até que série a mãe, pai ou responsável estudou (maior formação) 1 – Nunca estudou; 2 – Ensino fundamental cursando ou completo; 3 – Ensino médio completo; 4 – Ensino superior completo

Fonte: Barros (2016).

Com base no que foi exposto e na capacidade da TRI poder relacionar a medida do Inse com os itens utilizados na sua criação, permitindo calcular a probabilidade de um aluno responder determinada categoria dado o seu nível socioeconômico familiar, optou-se por analisar os agrupamentos hierárquicos com base no método de Ward (1963), para segmentar as escalas. Essa técnica, segundo Hair et. al. (2005), é utilizada para classificar indivíduos em grupos de modo que cada um seja muito semelhante aos indivíduos do grupo a que pertence e diferente em relação aos indivíduos dos demais grupos. Para a análise, utilizou-se o *software* IBM Spss Statistics, na versão 23.

A medida proposta consiste em utilizar apenas as informações que realmente são importantes para segmentação, que são os valores de Inse divididos em pequenas faixas e as probabilidades de resposta por item/categorias do modelo. Com isso, a segmentação não dependerá mais do banco de dados utilizado e fica relacionada apenas às informações usadas na modelagem, o que acaba por facilitar, também, a criação de série histórica entre as edições do indicador.

Com o intuito de definir o número de agrupamentos, foi considerado o coeficiente de aglomeração e o gráfico de dendrograma. O coeficiente de aglomeração é utilizado para ajudar a identificar grandes aumentos relativos na homogeneidade dos agrupamentos, e o gráfico dendrograma,

para explicitar o nível de similaridade entre os valores de Inse (HAIR et. al., 2005). A ideia é que essa metodologia forneça direcionamentos que facilitem a formação de grupos; porém, a referência será sempre a relação entre os itens e a medida.

Após a formação dos grupos definidos pelas faixas de Inse, foram calculadas as médias das probabilidades referentes a cada item/categoria em cada grupo, e, a partir dessa média, a categoria foi descrita na faixa em que ela apresentava a maior probabilidade. Optou-se por essa metodologia por considerar que, dentro do grupo, as probabilidades seriam bastante parecidas e evitaria-se, assim, a utilização de apenas um ponto para representar o todo.

Por fim, com base nas probabilidades médias, os grupos foram descritos de forma que cada um equivalesse a um nível, explicitando-se, assim, a ideia de ordem e hierarquia. Para a descrição da escala e posicionamento dos itens/categorias, foram utilizados três critérios, tal como descrito a seguir.

- Critério 1 – Probabilidades maiores ou iguais a 0,5.
- Critério 2 – Probabilidades maiores que 0,3 quando a probabilidade de alguma outra categoria do item estiver entre 0,5 e 0,6.
- Critério 3 – Quando nenhuma das situações anteriores ocorrer, verificar se a soma das probabilidades na junção de algumas categorias do item foi maior que 0,6.

Resultados

A escala proposta possui média 500 e desvio-padrão 100. Para aplicar a análise de agrupamentos hierárquicos, foram utilizados como variáveis de entrada os valores de Inse (de 100 a 900, com intervalos de 10 unidades, ou seja, 1/10 do desvio-padrão) e as probabilidades de resposta por item/categorias do modelo. Optou-se por utilizar intervalos de tamanho 10 a fim de conseguir separar de forma mais eficiente os grupos, para que a segmentação pudesse ser realizada. A seguir, na tabela 1, apresenta-se a estrutura da base de dados utilizada para a segmentação.

Tabela 1 Estrutura da base de dados para a segmentação

	Probabilidade calculada pelo modelo									
Inse	**Q02_B1***	**Q02_B2**	**Q02_B3**	**Q02_B4**	**Q03_B1**	**...**	**Q25_B1**	**Q25_B2**	**Q25_B3**	**Q25_B4**
100	1,00	0,00	0,00	0,00	0,80	...	0,41	0,56	0,02	0,00
110	1,00	0,00	0,00	0,00	0,79	...	0,39	0,59	0,02	0,00
120	1,00	0,00	0,00	0,00	0,77	...	0,36	0,61	0,02	0,00
...	...	...	...	...	...	...	...	...	...	...
880	0,00	0,11	0,21	0,67	0,00	...	0,00	0,01	0,06	0,92
890	0,00	0,10	0,19	0,71	0,00	...	0,00	0,01	0,06	0,93
900	0,00	0,08	0,17	0,74	0,00	...	0,00	0,01	0,05	0,94

*Questão 2, para grau de dificuldade da categoria k=1 do modelo politômico da TRI.

Com base nos resultados dessa análise, utilizou-se o coeficiente de aglomeração como critério para selecionar o número de grupos (HAIR et. al., 2005) e o dendrograma para representar o procedimento de agrupamento no qual uma medida de distância é estabelecida para formação dos grupos (valores de Inse) (BARROS, 2016). As soluções de quatro, seis e sete agrupamentos foram consideradas para a interpretação dos resultados.

Quadro 2 Opções de agrupamentos por meio de métodos hierárquicos

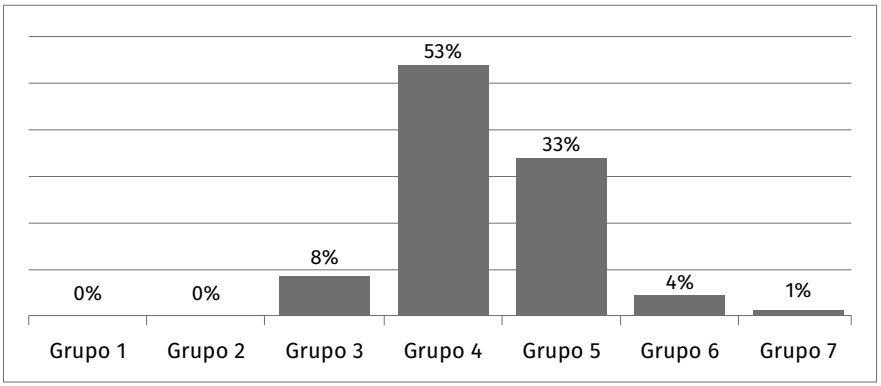
Opção com 4 Grupos	Opção com 6 Grupos	Opção com 7 Grupos
1 – De 100 a 380	1 – De 100 a 180	1 – De 100 a 180
	2 – De 190 a 380	2 – De 190 a 260
		3 – De 270 a 380
2 – De 390 a 520	3 – De 390 a 520	4 – De 390 a 520
3 – De 530 a 740	4 – De 530 a 650	5 – De 530 a 650
	5 – De 660 a 740	6 – De 660 a 740
4 – De 750 a 900	6 – De 750 a 900	7 – De 750 a 900

Fonte: Barros (2016).

Com base nesse resultado, optou-se por estudar primeiro a proposta mais abrangente, a de sete grupos. Em cada grupo, foi calculada a média das probabilidades nas faixas contempladas para cada item/categoria, e, segundo a probabilidade média, a categoria foi posicionada no grupo. Optou-se por essa metodologia porque, dentro do grupo, as probabilidades seriam bastante parecidas, o que evitaria, como ressaltado anteriormente, a utilização de apenas um ponto para representar o todo.

A escolha pela segmentação em sete grupos mostrou-se bastante apropriada no que se refere ao posicionamento de itens/categorias. Entretanto, observa-se que o grupo 4, que varia de 390 a 520, possui amplitude muito grande e, por incluir a média da distribuição, concentra um grande quantitativo de alunos. Dessa forma, seria interessante conseguir diferenciar melhor esses alunos de modo que um grande número deles não fique concentrado em um único grupo. Na figura 1, é apresentada a distribuição do Inse dos alunos considerando-se os sete grupos.

Figura 1 Distribuição do Inse dos alunos em sete grupos



Fonte: Barros (2016).

Com base no resultado apresentado, optou-se por aplicar o mesmo procedimento relativo aos agrupamentos hierárquicos, porém, considerando apenas os valores de Inse referentes aos grupos 4 e 5 separadamente e buscando uma proposta para diferenciar melhor os alunos pertencentes a esses dois grupos. Esse procedimento, não obstante, só é possível caso existam itens/categorias suficientes para separar esses dois grupos.

Na tabela 2, são apresentadas as probabilidades médias de respostas por item/categoria em cada grupo. Foram destacadas as células que obedeceram aos critérios estabelecidos ao final da proposta de segmentação. O critério 1 foi apresentado na cor cinza; o critério 2, na cor cinza-escuro; o critério 3, na cor branca; e, quando nenhum desses critérios foi satisfeito, a célula não foi destacada.

Tabela 2 Probabilidade média por itens/categorias nos grupos

	Grupo 1	Grupo 2	Grupo 3	Grupo 4_1	Grupo 4_2	Grupo 5_1	Grupo 5_2	Grupo 6	Grupo 7
Q02 – TV por assinatura em casa									
1 – Não tem	**1,00**	**1,00**	**0,98**	**0,93**	**0,79**	**0,48**	0,24	0,09	0,01
2 – Sim, uma	0,00	0,00	0,02	0,07	0,20	**0,48**	**0,66**	**0,65**	**0,27**
3 – Sim, duas	0,00	0,00	0,00	0,00	0,01	0,03	0,07	0,17	**0,27**
4 – Sim, três ou mais	0,00	0,00	0,00	0,00	0,00	0,01	0,03	0,09	**0,45**
Q03 – Videocassete e/ou DVD em casa									
1 – Não tem	**0,74**	**0,58**	**0,36**	0,21	0,13	0,07	0,04	0,02	0,01
2 – Sim, um	0,25	**0,41**	**0,59**	**0,70**	**0,71**	**0,65**	**0,54**	**0,41**	0,20
3 – Sim, dois	0,01	0,02	0,04	0,08	0,13	0,23	**0,32**	**0,40**	**0,41**
4 – Sim, três ou mais	0,00	0,00	0,01	0,02	0,03	0,06	0,10	0,17	**0,38**
Q05 – Internet em casa									
1 – Não tem	**0,99**	**0,97**	**0,81**	**0,48**	0,20	0,05	0,01	0,00	0,00
2 – Sim, uma	0,01	0,03	0,19	**0,52**	**0,78**	**0,87**	**0,77**	**0,48**	0,10
3 – Sim, duas	0,00	0,00	0,00	0,00	0,01	0,05	0,14	**0,25**	0,14
4 – Sim, três ou mais	0,00	0,00	0,00	0,00	0,00	0,03	0,09	**0,27**	**0,76**
Q06 – Rádio em casa									
1 – Não tem	**0,67**	**0,55**	**0,41**	0,29	0,22	0,14	0,10	0,07	0,04
2 – Sim, um	0,30	**0,40**	**0,51**	**0,58**	**0,61**	**0,59**	**0,55**	**0,49**	**0,35**
3 – Sim, dois	0,02	0,04	0,07	0,10	0,14	0,20	0,25	**0,30**	**0,37**
4 – Sim, três ou mais	0,01	0,01	0,02	0,03	0,04	0,07	0,09	0,14	**0,25**

	Grupo 1	Grupo 2	Grupo 3	Grupo 4_1	Grupo 4_2	Grupo 5_1	Grupo 5_2	Grupo 6	Grupo 7
Q08 – Telefone celular em casa									
1 – Não tem	**0,35**	0,19	0,08	0,03	0,02	0,01	0,00	0,00	0,00
2 – Sim, um	**0,57**	**0,65**	**0,57**	**0,41**	**0,26**	0,14	0,07	0,04	0,01
3 – Sim, dois	0,05	0,11	0,21	**0,27**	**0,27**	0,21	0,14	0,08	0,03
4 – Sim, três ou mais	0,03	0,06	0,15	**0,28**	**0,45**	**0,65**	**0,79**	**0,88**	**0,96**
Q09 – Aspirador de pó em casa									
1 – Não tem	**1,00**	**1,00**	**0,99**	**0,96**	**0,89**	**0,67**	**0,41**	0,19	0,04
2 – Sim, um ou mais	0,00	0,00	0,01	0,04	0,11	0,33	**0,59**	**0,81**	**0,96**
Q10 – Geladeira em casa									
1 – Não tem	0,34	0,17	0,07	0,03	0,01	0,00	0,00	0,00	0,00
2 – Sim, uma	**0,66**	**0,83**	**0,92**	**0,94**	**0,92**	**0,85**	**0,74**	**0,56**	0,26
3 – Sim, duas	0,00	0,00	0,01	0,03	0,06	0,13	0,23	**0,37**	**0,50**
4 – Sim, três ou mais	0,00	0,00	0,00	0,00	0,01	0,02	0,03	0,07	0,24
Q11 – Freezer (aparelho independente ou parte da geladeira) em casa									
1 – Não tem	**0,99**	**0,97**	**0,92**	**0,83**	**0,70**	**0,49**	0,33	0,19	0,07
2 – Sim, um ou mais	0,01	0,03	0,08	0,17	0,30	**0,51**	**0,67**	**0,81**	**0,94**
Q12 – Máquina de lavar roupa em casa									
1 – Não tem	**0,98**	**0,92**	**0,70**	0,38	0,17	0,05	0,02	0,01	0,00
2 – Sim, uma	0,02	0,08	0,30	**0,61**	**0,82**	**0,92**	**0,90**	**0,76**	0,32
3 – Sim, duas ou mais	0,00	0,00	0,00	0,00	0,01	0,03	0,09	0,24	**0,68**
Q13 – Automóvel em casa									
1 – Não tem	**0,99**	**0,98**	**0,90**	**0,73**	**0,50**	0,22	0,10	0,04	0,01
2 – Sim, um	0,01	0,02	0,09	0,25	**0,44**	**0,59**	**0,53**	**0,34**	0,10
3 – Sim, dois	0,00	0,00	0,01	0,02	0,05	0,16	0,29	**0,41**	**0,30**
4 – Sim, três ou mais	0,00	0,00	0,00	0,00	0,01	0,03	0,08	0,20	**0,59**

	Grupo 1	Grupo 2	Grupo 3	Grupo 4_1	Grupo 4_2	Grupo 5_1	Grupo 5_2	Grupo 6	Grupo 7
Q14 – Banheiro em casa									
1 – Não tem	**0,67**	0,31	0,07	0,01	0,00	0,00	0,00	0,00	0,00
2 – Sim, um	0,33	**0,69**	**0,91**	**0,92**	**0,76**	**0,41**	0,17	0,05	0,01
3 – Sim, dois	0,00	0,00	0,02	0,06	0,21	**0,48**	**0,54**	0,34	0,07
4 – Sim, três ou mais	0,00	0,00	0,00	0,01	0,03	0,11	0,29	**0,60**	**0,93**
Q15 – Quartos para dormir em casa									
1 – Não tem	0,20	0,09	0,03	0,01	0,01	0,00	0,00	0,00	0,00
2 – Sim, um	**0,58**	**0,50**	0,29	0,15	0,08	0,03	0,02	0,01	0,00
3 – Sim, dois	0,20	**0,36**	**0,53**	**0,55**	**0,45**	0,28	0,17	0,09	0,03
4 – Sim, três	0,02	0,05	0,13	0,25	**0,39**	**0,51**	**0,53**	**0,44**	0,21
5 – Sim, quatro ou mais	0,00	0,01	0,02	0,04	0,08	0,17	0,29	**0,46**	**0,76**
Q19 – Contrata empregado/a doméstico/a em casa									
1 – Não	**1,00**	**1,00**	**1,00**	**1,00**	**0,99**	**0,95**	**0,79**	**0,42**	0,05
2 – Sim, um ou mais	0,00	0,00	0,00	0,00	0,01	0,06	0,21	**0,58**	**0,95**
Q20 – Renda mensal da família (considerando a renda do aluno)									
1 – Nenhuma renda	**0,89**	**0,51**	0,09	0,01	0,00	0,00	0,00	0,00	0,00
2 – Até um S.M.	0,11	**0,48**	**0,80**	**0,54**	0,18	0,02	0,00	0,00	0,00
3 – Mais de um até 1,5 S.M.	0,00	0,01	0,08	**0,30**	**0,30**	0,08	0,01	0,00	0,00
4 – Mais de 1,5 até 3 S.M.	0,00	0,00	0,02	0,14	**0,41**	**0,39**	0,13	0,03	0,00
5 – Mais de 3 até 7 S.M.	0,00	0,00	0,00	0,02	0,10	**0,40**	**0,48**	**0,21**	0,02
6 – Mais de 7 até 9 S.M.	0,00	0,00	0,00	0,00	0,01	0,05	**0,16**	**0,15**	0,02
7 – Mais de 9 até 12 S.M.	0,00	0,00	0,00	0,00	0,00	0,03	0,11	**0,19**	0,05
8 – Acima de 12 S.M.	0,00	0,00	0,00	0,00	0,00	0,02	0,11	**0,42**	**0,91**

	Grupo 1	Grupo 2	Grupo 3	Grupo 4_1	Grupo 4_2	Grupo 5_1	Grupo 5_2	Grupo 6	Grupo 7
Q21– Mãe, ou a mulher responsável por você, sabe ler e escrever									
1 – Não	**0,72**	**0,46**	0,19	0,07	0,03	0,01	0,00	0,00	0,00
2 – Sim	0,28	**0,54**	**0,81**	**0,93**	**0,97**	**0,99**	**1,00**	**1,00**	**1,00**
Q22– Pai, ou homem responsável por você, sabe ler e escrever									
1 – Não	**0,73**	**0,52**	0,27	0,12	0,06	0,03	0,01	0,01	0,00
2 – Sim	0,27	**0,48**	**0,73**	**0,88**	**0,94**	**0,97**	**0,99**	**0,99**	**1,00**
Q25 – Até que série a mãe, pai ou responsável estudou (maior formação)									
1 – Nunca estudou	0,32	0,17	0,07	0,03	0,01	0,01	0,00	0,00	0,00
2 – Ensino Fundamental cursando ou completo	**0,64**	**0,76**	**0,73**	**0,60**	**0,44**	0,26	0,15	0,08	0,03
3 – Ensino Médio Completo	0,03	0,07	0,16	0,28	**0,37**	**0,42**	**0,37**	0,27	0,11
4 – Ensino Superior Completo	0,01	0,01	0,04	0,09	0,17	**0,32**	**0,48**	**0,65**	**0,86**

Fonte: Barros (2016).

Observa-se que os dois grupos que foram segmentados novamente possuem alguns itens/categorias que permitiram a diferenciação dos alunos, viabilizando, assim, a melhoria da segmentação da escala. No entanto, para que tal processo fosse mais apurado, seria interessante incluir mais itens que pudessem diferenciar os alunos nas faixas de nível socioeconômico entre 390 e 650, correspondentes aos grupos 4 e 5, para, dessa maneira, caracterizar melhor a distinção entre esses grupos.

Os grupos tiveram sua escala descrita utilizando-se itens/categorias cujas células foram destacadas em cinza, cinza-escuro e branco. Ao apresentar a escala, algumas categorias foram omitidas para simplificar a escrita, mas sem comprometê-la. Por exemplo, no primeiro grupo, o de menor nível socioeconômico, as famílias, em geral, não possuem renda e, por essa razão, apesar de apresentarem probabilidades mais altas de não contratarem serviço de empregado(a) mensalista e de não terem automóvel em casa, é desnecessário explicitar a baixa probabilidade de contratarem

esse serviço ou de possuírem esse bem. Da mesma forma, no último grupo, que é caracterizado pelos alunos de nível socioeconômico mais alto, é desnecessário dizer que os pais ou responsáveis sabem ler e escrever.

Outra alteração realizada ao descrever a escala, conforme ressaltado na sessão referente ao método, foi a troca do termo "grupos" por "níveis". Isso porque a palavra grupo não sugere a ideia de ordenamento e hierarquia, ao passo que a palavra nível, ao explicitar essas ideias, se mostra mais coerente com o construto do indicador proposto e o modelo da TRI adotado. Não obstante, é importante destacar que esses nomes apenas correspondem a uma representação ordinal, o que indica que, na escala do Inse, o nível 1 é o mais baixo, e o nível 9, o mais alto. Além disso, é preciso advertir que não é possível, devido aos nomes atribuídos aos níveis, fazer equiparações com outras escalas, pois o significado de cada um deles decorre, necessariamente, do referencial teórico e metodológico adotado, o que significa dizer que os agrupamentos obtidos por cada referencial tendem a ser diferentes em relação aos demais. Para conhecer os níveis, é preciso remeter à descrição da escala, apresentada no quadro 3.

Quadro 3 Descrição dos níveis

Nível 1 (grupo 1) – Inse de 100 a 180
Esse grupo é composto pelos alunos com o menor nível socioeconômico familiar. De modo geral, esses alunos indicaram que a família não possui renda familiar. No que se refere à escolaridade, eles informaram que os pais ou responsáveis não sabem ler e escrever e que estes possuem ensino fundamental completo ou incompleto. A casa que moram contém, em geral, um quarto para dormir, não possui banheiro e algumas famílias têm um telefone celular e uma geladeira.
Nível 2 (grupo 2) – Inse de 180 a 260
Nesse grupo algumas famílias não possuem renda e outras recebem até 1 salário mínimo. Boa parte dos pais ou responsáveis sabe ler e escrever e possui o ensino fundamental completo ou incompleto. A casa que moram contém, em geral, um ou dois quartos para dormir e um banheiro. A família possui um telefone celular, uma geladeira e, em alguns casos, um rádio e um videocassete e/ou DVD.

Nível 3 (grupo 3) – Inse de 260 a 380

Em geral, a família possui renda familiar de até 1 salário mínimo. Os pais ou responsáveis sabem ler e escrever e têm ensino fundamental completo ou incompleto.
A casa que moram contém, em geral, dois quartos para dormir e um banheiro. A família possui um telefone celular, uma geladeira e, em alguns casos, um rádio e um videocassete e/ou DVD.

Nível 4 (grupo 4_1) – Inse de 380 a 430

Em geral, a família possui renda familiar de até 1,5 salário mínimo. Os pais ou responsáveis sabem ler e escrever e têm ensino fundamental completo ou incompleto.
A casa contém, em geral, dois quartos para dormir e um banheiro. A família possui em casa um rádio, um videocassete e/ou DVD, uma máquina de lavar roupa, uma geladeira e mais de um telefone celular. Além disso, algumas famílias têm internet em casa.

Nível 5 (grupo 4_2) – Inse de 430 a 520

A renda mensal da família é de 1 a 3 salários mínimos. Os pais ou os responsáveis sabem ler e escrever e possuem ensino fundamental incompleto ou já concluíram o ensino fundamental ou o ensino médio.
A casa contém, em geral, dois ou três quartos para dormir e um banheiro. A família possui em casa um rádio, um videocassete e/ou DVD, internet, uma máquina de lavar roupa, uma geladeira e mais de um telefone celular. Além disso, algumas famílias possuem um automóvel.

Nível 6 (grupo 5_1) – Inse de 520 a 600

A renda mensal da família é de 1,5 a 7 salários mínimos. Os pais possuem ensino médio ou ensino superior completo.
A casa contém, em geral, três quartos para dormir e um ou dois banheiros. A família possui três ou mais telefones celulares, um rádio, um videocassete e/ou DVD, internet, um automóvel, uma máquina de lavar roupa e uma geladeira. Algumas famílias possuem alguns bens e outras não, como: *freezer* (aparelho independente ou parte da geladeira) e TV por assinatura.

Nível 7 (grupo 5_2) – Inse de 600 a 650

A renda mensal da maior parte das famílias pertencentes a esse nível é de 3 a 7 salários mínimos, e da menor parte, de 7 a 9 salários mínimos. Os pais possuem ensino médio ou superior completos.
A casa contém, em geral, três quartos para dormir e dois banheiros. A família possui três ou mais telefones celulares, um rádio, um ou dois videocassetes e/ou DVD, TV por assinatura, internet, um automóvel, uma máquina de lavar roupa e uma geladeira, *freezer* (aparelho independente ou parte da geladeira). Algumas famílias possuem aspirador de pó.

Nível 8 (grupo 6) – Inse de 650 a 740

A renda mensal das famílias é bastante variada. Algumas têm renda de 3 a 7 salários mínimos, umas de 7 a 12 e outras acima de 12. Os pais ou responsáveis possuem ensino superior completo.

A casa contém, em geral, três ou mais quartos para dormir e três ou mais banheiros. A família possui três ou mais telefones celulares, um ou dois rádios, um ou dois videocassetes e/ou DVD, TV por assinatura, internet, uma máquina de lavar roupa, uma ou duas geladeiras, *freezer* (aparelho independente ou parte da geladeira), aspirador de pó e um ou dois automóveis. Algumas famílias contratam o serviço de empregado(a) doméstico(a).

Nível 9 (grupo 7) – Inse de 740 a 900

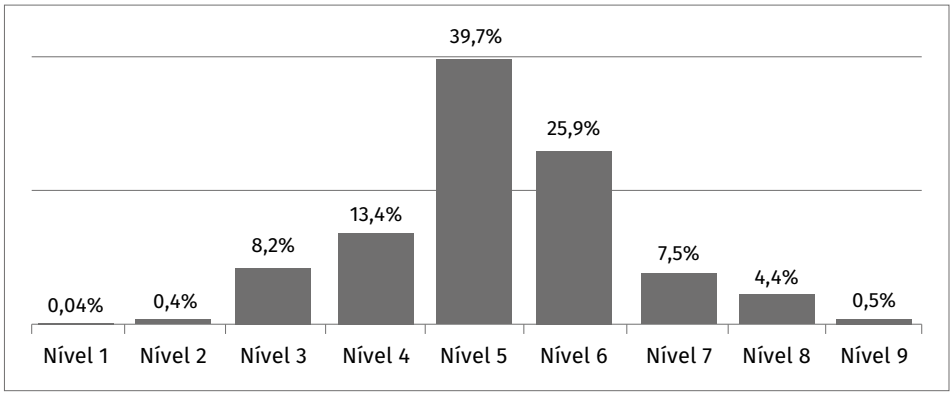
A renda mensal da família é maior que 12 salários mínimos. Os pais ou responsáveis possuem ensino superior completo.

A casa contém, em geral, quatro ou mais quartos para dormir e três ou mais banheiros. A família possui três ou mais telefones celulares, mais de um rádio, dois ou mais videocassetes e/ou DVD, TV por assinatura, internet, duas ou mais máquinas de lavar roupa, duas geladeiras, *freezer* (aparelho independente ou parte da geladeira), aspirador de pó e dois ou mais automóveis. Todas as famílias contratam o serviço de empregada doméstica.

Fonte: Barros (2016).

Na figura 2, é apresentada a distribuição do Inse dos alunos estudados, com base na definição dos nove níveis.

Figura 2 Distribuição por Inse dos alunos em nove níveis



Fonte: Barros (2016).

Ao analisar a figura 2 e o quadro 3, verifica-se que a escala do indicador é sensível para captar camadas da sociedade que possuem um nível socioeconômico muito baixo, isto é, aquelas que declararam que a família não possui renda familiar, que os pais ou responsáveis não sabem ler e escrever e que possuem ensino fundamental completo ou incompleto. O indicador também é sensível no que se refere às camadas da população com nível socioeconômico mais alto, diferenciando, por exemplo, os alunos com renda mensal familiar maior que 12 salários mínimos, cujos pais ou responsáveis possuem ensino superior completo, em relação aos demais.

Esses resultados evidenciam a vantagem de segmentar a escala utilizando os parâmetros do modelo, tanto porque se deixa de depender da base de dados existente, quanto porque se informa de maneira mais clara e precisa o que o indicador está realmente mensurando, ao relacionar os níveis com as descrições de itens/categorias nas diversas partes da escala.

Discussão

Esse trabalho é uma proposta de segmentação de grupos e construção de escala a partir de indicadores elaborados utilizando-se a TRI. Usou-se o Inse (BARROS, 2016), pois este está cada vez mais presente nas discussões no âmbito da educação, já que permite uma visão mais contextualizada da realidade em que o aluno está inserido e subsidia, com mais informações, a formulação e a implementação de políticas educacionais.

Buscou-se também estudar como os itens se relacionam com a medida criada neste trabalho. Para tanto, estimou-se o Inse de cada aluno e, a partir das relações entre os parâmetros dos itens e do construto nível socioeconômico, buscou-se a melhor forma de agrupar os alunos em faixas de nível socioeconômico, de modo que eles fossem semelhantes dentro das faixas e diferentes entre as faixas, constituindo-se, com a adoção dessa técnica, nove grupos de nível socioeconômico distintos.

Por fim, cada grupo recebeu um nível e foi descrito a partir dos itens/categorias que melhor designavam suas características. A possibilidade de relacionar a medida criada com os itens que definem o indicador é um dos grandes

diferenciais da TRI quando comparada com outras técnicas, pois contribui efetivamente para a compreensão e a comunicação do que a medida realmente expressa.

Diante do exposto, é preciso destacar ainda que no Inse criado pelo Inep (2014d), a segmentação foi usada somente para formação de grupos das escolas e não diretamente dos alunos. Para a interpretação da escala, o Inep (2014d) utilizou a metodologia proposta por Huynh (1998), que considera intervalos equidistantes e não busca um ponto ótimo para a segmentação da escala.

Por outro lado, a despeito da contribuição da proposta de segmentação de escala apresentada, acredita-se que trabalhos futuros possam ser desenvolvidos buscando-se empregar ou criar outras técnicas de segmentação e descrição de escala para modelos politômicos da TRI, capazes de aprimorar a inteligibilidade, a precisão e a comunicação da medida. Do mesmo modo, comparar a eficiência dos métodos existentes na literatura aparece, nesse cenário, como outra tarefa importante a ser desenvolvida pelos pesquisadores da área.

Referências

ALVES, M. T. G.; SOARES, J. F. Medidas de nível socioeconômico em pesquisas sociais: uma aplicação aos dados de uma pesquisa educacional. *Opinião Pública,* Campinas, vol. 15, n. 1, p. 1-30, 2009.

ALVES, M. T. G.; SOARES, J. F. Contexto escolar e indicadores educacionais: condições desiguais para a efetivação de uma política de avaliação educacional. *Educação e Pesquisa*, São Paulo, v. 39, n. 1, p. 177-194, 2013.

ALVES, M. T. G.; SOARES, J. F.; XAVIER, F. P. Índice socioeconômico das escolas de educação básica brasileiras. *Ensaio: Avaliação e Políticas Públicas em Educação*, Rio de Janeiro, v. 22, n. 84, p. 671-703, 2014.

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

BARROS, G. T. de. F. *Procedimentos para a construção de indicadores por meio da Teoria de Resposta ao Item*: a criação de uma medida de nível socioeconômico familiar. Florianópolis: Centro Tecnológico da Universidade Federal de Santa Catarina, 2016. (Dissertação de mestrado profissional pelo Programa de Pós-Graduação em Métodos e Gestão em Avaliação).

BEATON, A. E.; ALLEN, N. L. Interpreting Scales through Scale Anchoring. *Journal of Educational Statistics*, v. 17, p. 191-204, 1992.

BORTOLOTTI, S.; VINCENZI L. et al. Avaliação do nível de satisfação de alunos de uma instituição de ensino superior: uma aplicação da teoria da resposta ao item. *Gestão e Produção*, São Carlos, v. 19, n. 2, p. 287-302, 2012.

BRASIL. Lei nº 13.005, de 25 de junho de 2014. Aprova o Plano Nacional de Educação e dá outras providências. *Diário Oficial da União,* Brasília, DF, 26 jun. 2014. Seção 1, p. 1. Edição Extra. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l13005.htm>.

BUCHMANN, C. Measuring Family Background in International Studies of Education: Conceptual Issues and Methodological Challenges. In: PORTER, A. e GAMORAN, A. (Ed.). *Methodological Advances in Cross-National Surveys of Educational Achievement*. Washington, DC: National Academy Press, p.150-197, 2002.

CIZEK, G. J. *Setting Performance Standards*: Theory and Applications. New York: Routledge, 2001.

FRANCO, C. Quais as contribuições da avaliação para as políticas educacionais? In: BONAMINO, A; BESSA, N; FRANCO, C. *Avaliação da educação básica*. Rio de Janeiro: Editora PUC-Rio; São Paulo: Loyola, 2004.

HAIR JR., J. F. et al. *Análise multivariada de dados*. 5. ed. Porto Alegre: Bookman, 2005.

HUYNH, H. On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics*. v. 23, n. 1, p. 35-56, 1998.

IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY: IBM Corp.

INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira). Indicador de adequação da formação do docente na educação básica. Brasília: Inep, 2014a.

INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira). Indicador para mensurar a complexidade da gestão nas escolas a partir dos dados do Censo Escolar da Educação Básica. Brasília: Inep, 2014b.

INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira). Indicador de Esforço Docente. Brasília: Inep, 2014c.

INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira). Indicador de Nível Socioeconômico das Escolas de Educação Básica (Inse) participantes da Avaliação Nacional da Alfabetização (ANA). Brasília: Inep, 2014d.

INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira). Indicador de regularidade do docente da Educação Básica. Brasília: Inep, 2015.

JANNUZZI, P. M. *Indicadores sociais no Brasil*: conceitos, fontes e aplicações. Campinas: Alínea/PUC Campinas, 2001.

KLEIN, R. Utilização da Teoria de Resposta ao Item no Sistema Nacional de Avaliação da Educação Básica (Saeb). *Revista Meta*: Avaliação, v. 1, n. 2, p. 125-140, 2009.

NETO, J. J. S. et al. Uma escala para medir a infraestrutura escolar. *Est. Aval. Educ.*, São Paulo, v. 24, n. 54, p. 78-99, 2013.

SAMEJIMA F. A. *Estimation of latent ability using a response pattern of graded scores*. Psychometric Monograph, n. 17, 1969.

SOARES, T. M. Utilização da Teoria de Resposta ao Item na Produção de Indicadores Sócio-Econômicos. *Pesquisa Operacional*, v. 25, n. 1, p. 83-112, 2005.

SOARES, T. M.; SOUZA, R. C.; PEREIRA, V. R. Métodos Alternativos ao Critério Brasil para Construção de Indicadores Sócio-Econômicos: Teoria da Resposta ao Item. In: XXXVI SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL, 2004, São João Del Rei. Anais, Rio de Janeiro: Sobrapo, 2004.

WARD, J. H. Hierarquical grouping to optimize an objective function. *Journal of the American Statistical Association*, v. 58, p. 236-244, 1963.

M.ª Gabriela Thamara de Freitas Barros
Mestre em Métodos e Gestão em Avaliação pela Universidade Federal de Santa Catarina
Pesquisadora do Inep
gabybarross@gmail.com

Adriano Ferreti Borgatto
Doutor em Agronomia pela Escola Superior de Agricultura Luiz Queiroz
Professor da Universidade Federal de Santa Catarina
borgatto@inf.ufsc.br

Adolfo Samuel de Oliveira
Doutor em Educação pela Universidade de São Paulo
Pesquisador do Inep
adolfoso77@yahoo.com.br

ESTUDOS BRASILEIROS SOBRE EFICÁCIA ESCOLAR: UMA REVISÃO DE LITERATURA

BRAZILIAN STUDIES ON SCHOOL EFFECTIVENESS: A LITERATURE REVIEW

ESTUDIOS BRASILEÑOS SOBRE LA EFICACIA ESCOLAR: UNA REVISIÓN DE LITERATURA

Camila Akemi Karino

Jacob Arie Laros

RESUMO

Na busca por uma educação de qualidade, vários estudos foram iniciados na década de 1960 a fim de avaliar se as escolas contribuem significativamente para a formação dos estudantes, o que levou ao surgimento da área de eficácia escolar. Neste estudo, tem-se como objetivo principal realizar uma revisão sistemática da literatura brasileira na área de eficácia escolar. Foram analisados 30 artigos publicados em periódicos científicos entre 2000 e 2013. Foi possível identificar efeitos consistentes de fatores como nível socioeconômico, etnia e atraso escolar sobre o desempenho escolar, bem como ficaram evidentes dificuldades de definição e mensuração de fatores intraescolares. Ressalta-se também que problemas de igualdade e equidade precisam ser melhor investigados. Os resultados demonstram similaridades com a literatura estrangeira e ajudam a nortear estudos futuros na área de eficácia escolar.

Palavras-chaves: eficácia escolar; desempenho escolar; avaliação escolar.

ABSTRACT

In search for quality education, several studies were initiated in the decade of 1960 with the objective to evaluate to what extent schools contribute to the formation of their students. These studies have led to the emergence of the area of school effectiveness. The main objective of this study was to accomplish a systematic review of the Brazilian literature in the area of school effectiveness. Thirty articles published between 2000 and 2013 in Brazilian scientific journals were analyzed. Solid effects of factors like socioeconomic level, race and school delay on school

performance were identified. In addition, difficulties related to the definition and measurement of intra-school factors became evident. It should be emphasized that problems of equality and equity need to be investigated better. The results showed similarities with the results of the scientific literature of other countries and can guide future studies in the area of school effectiveness.

Keywords: school effectiveness; school performance; educational assessment.

RESUMEN

En la búsqueda de una educación de calidad, diversos estudios fueron iniciados en la década de los 60 con la finalidad de verificar si las escuelas contribuyen en la formación de los estudiantes, lo que llevó al surgimiento del área de eficacia escolar. En esta investigación, el objetivo central es realizar una revisión sistemática de la literatura brasileña en el campo de la eficacia escolar. Fueron analizados 30 artículos publicados en periódicos científicos entre los años 2000 y 2013. Fue posible identificar efectos consistentes de aspectos como nivel socioeconómico, etnia y atraso escolar sobre el desempeño escolar. También fueron evidentes los efectos de factores relativos a la escuela. Por fin se destaca que problemas de igualdad y equidad necesitan ser major investigados. Los resultados muestran semejanza con la literatura extranjera y ayudan a direccionar estudios futuros en el área de eficacia escolar.

Palavras clave: eficacia escolar; desempeño escolar; evaluación escolar.

Introdução

É fascinante o enigma que envolve descobrir porque algumas escolas têm maior capacidade do que outras de proporcionar melhores resultados acadêmicos (LEE, 2008). Essa afirmação provoca reflexão sobre a importância dos contextos nos quais os alunos se desenvolvem para o progresso educacional. Isso porque, além da potencialidade individual e de outros aspectos que o aluno traz consigo, fatores contextuais exercem uma forte influência no aprendizado.

Tendo como meta a promoção de uma educação de qualidade, vários estudos foram iniciados na década de 1960, a fim de avaliar se as escolas contribuem significativamente para a formação dos estudantes. Em específico,

buscava-se identificar quais fatores escolares colaboram para uma maior eficácia escolar – termo este utilizado para se referir à capacidade de a escola contribuir para que seus alunos alcancem resultados para além dos esperados, considerando características contextuais e escolares (MORTIMORE, 1991). Para Reymolds et al. (2000), a área de eficácia escolar busca compreender e analisar o papel da escola por meio de estudos quantitativos e qualitativos sobre efeito-escola (*school effects)*, escolas efetivas ou eficazes (*effective schools*) e formas de melhoria da escola (*school improvement*).

Nesses quase 50 anos de pesquisas na área de eficácia escolar, muito já foi produzido e discutido. Neste artigo, tem-se como objetivo principal realizar uma análise da produção brasileira nessa área, buscando contrapô-la à literatura estrangeira, delineando lacunas que permitam subsidiar uma agenda de pesquisa.

Histórico dos estudos sobre eficácia escolar

O relatório publicado por Coleman et al. (1966) é tido como um marco na área de eficácia escolar. Nele, os pesquisadores afirmaram que cerca de 90% da variação no desempenho acadêmico é explicada pelas condições socioeconômicas dos alunos e de suas famílias e que diferenças entre as escolas representam uma interferência muito pequena no desempenho acadêmico. Dessa forma, o relatório impactou toda a comunidade acadêmica ao apontar que escolas têm pouca influência no desempenho de um aluno quando controlados o contexto socioeconômico e os antecedentes sociais.

O impacto gerado por esses resultados impulsionou uma série de outros estudos que buscaram compreender se as escolas realmente podem fazer a diferença (EDMONDS, 1979; MORTIMORE et al. 1988; RUTTER et al. 1979). Essa reação delineou a constituição da área de eficácia escolar e contribuiu para o avanço significativo nas discussões sobre equidade e igualdade educacional, bem como no entendimento de como fatores escolares e extraescolares interferem no desempenho acadêmico.

Ao surgir, em reação aos resultados apontados pelo Relatório Coleman, a área de eficácia escolar desenvolve-se inicialmente em um clima reativo, muito focado no papel da escola e na busca incessante por encontrar

fatores que expliquem a eficácia educacional. Essas características iniciais exerceram forte influência na produção da área tanto em termos teóricos quanto metodológicos.

Em termos teóricos, ainda no final do século XX, os estudos sobre eficácia escolar são criticados por consistirem em numerosos estudos empíricos, carentes de uma teoria sólida que pudesse ser verificada (VAN DEN EEDEN; HOX; HAUER, 1990). Apenas mais recentemente é que se tem constituído uma teoria da área que visa explicar a associação e a dinâmica entre as variáveis (REYNOLDS et al. 2011).

Ao analisar milhares de estudos a partir de uma meta-análise, Teodorovic (2009) identificou que a maioria dos estudos ao longo do desenvolvimento da área pode ser categorizada em três grupos, conforme seus paradigmas teóricos: *i*) os estudos de *input-output*; *ii*) os estudos de eficácia escolar e *iii*) os estudos de eficácia docente.

A fundamentação teórica dos estudos de *input-output* é a de que os *inputs* da escola (recursos humanos, infraestrutura, verba) são os determinantes para um bom desempenho. Essa concepção norteou os estudos iniciais na área de monitoramento de sistema e, de forma geral, buscava demonstrar o efeito de fatores econômicos, organizacionais e de estrutura sobre os resultados da escola. Carvallo-Pontón (2010) complementa que esses foram os primeiros estudos na área, nos quais se utilizavam técnicas de regressão linear e, muitas vezes, havia alta colinearidade entre as variáveis.

Os resultados dos estudos de *input-output*, em geral, foram contraditórios e sugeriam efeitos fracos dos *inputs*. Para Willms (1992), vários fatores contribuíram para esses resultados: *i*) ausência de grandes diferenças nas políticas e práticas adotadas entre as escolas; *ii*) ausência de especificação de efeitos entre níveis; *iii*) limitação dos estudos a fatores fáceis de serem medidos, uma vez que os principais processos que afetam a escolarização são de difícil definição e mensuração; e *iv*) não abrangência de fatores contextuais nos estudos.

Além disso, os estudos de *input-output* foram criticados por não apresentarem fatores que pudessem ajudar os educadores a melhorarem

as escolas. Nesses estudos, as escolas eram tratadas como "caixas pretas", investigando-se apenas os insumos e os produtos. Na tentativa de desvendar as "caixas pretas", surgem os estudos de eficácia escolar propriamente ditos.

A diferença em relação aos estudos anteriores é o foco nos processos escolares que ocorriam dentro da "caixa-preta" da escola (ALVES; SOARES, 2007a). Esses estudos de eficácia escolar concebem que o desempenho do aluno é determinado por fatores escolares e processuais, tais como: presença de um bom líder, foco nos objetivos a serem alcançados, altas expectativas, cooperação, clima acadêmico e monitoramento constante (TEODOROVIC, 2009).

Os estudos de eficácia escolar passam a utilizar um modelo aprimorado do antigo *input-output*, uma vez que reconhecem a estrutura multinível e buscam separar recursos de entrada dos fatores processuais (WILLMS, 1992). O objetivo era estimar a magnitude dos efeitos escolares, entendidos como a capacidade de as escolas influenciarem o desenvolvimento do aluno, além de trazer mais informações processuais que permitissem melhor compreender as razões de algumas escolas terem melhores resultados do que outras (CARVALLO-PONTÓN, 2010). Engloba-se, portanto, estudos que buscam estimar o efeito-escola (nas suas várias dimensões: valor agregado, eficácia diferencial, equidade social), identificar fatores associados ao desempenho, identificar características que tornam umas escolas mais eficazes do que outras e encontrar meios de promoção de mudanças na escola.

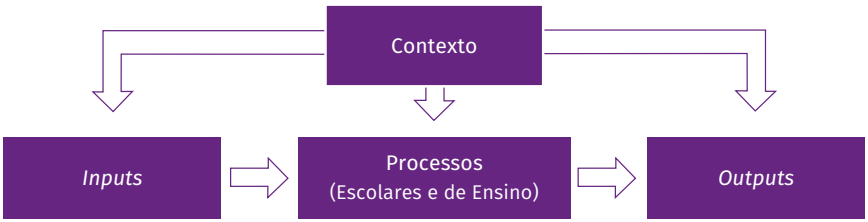
Nesse contexto de compreender melhor as dinâmicas intraescolares, surgem também as pesquisas que buscam explorar o efeito de diferentes tipos de práticas de ensino sobre o desempenho acadêmico. Esses estudos concentraram-se em verificar os comportamentos dos professores e as práticas de sala de aula utilizadas para explicar o alcance de um bom desempenho (TEODOROVIC, 2009).

Se por um lado os estudos de eficácia docente (*instructional effectiveness*) podem ser classificados como pertencentes a um paradigma específico, por outro as características de ensino podem também ser utilizadas como fatores processuais nos estudos de eficácia escolar. A diferença mais

marcante quando características processuais de ensino são incorporadas aos modelos de eficácia escolar é o acréscimo de outro nível de análise relativo ao professor ou à turma.

Ao sistematizar a produção na área, Scheerens (1990) propõe o modelo básico de funcionamento da escola. Esse modelo busca integrar e ser uma estrutura de referência das variáveis que exercem efeito sobre o desempenho escolar (*output*), categorizando-as em contextuais, de *input* e de processo. A figura 1 representa a proposta do modelo integrado.

Figura 1 Modelo integrado de avaliação da eficácia escolar



Fonte: Scheerens, 1990.

No modelo apresentado, os fatores contextuais são aqueles fatores extraescolares que impactam o processo educativo, mas que a escola possui pouco ou nenhum gerenciamento sobre eles. Scheerens (1990) menciona, por exemplo, gestão administrativa, localização da escola (rural ou urbana) e tamanho da escola. Soares (2007) cita outros possíveis fatores extraescolares: valores sociais, políticas públicas, recursos, gestão do sistema pelas secretarias educacionais e características socioeconômicas e culturais da comunidade.

A segunda parte do modelo integrado corresponde às ações intraescolares, que envolvem: *input*, processo e *output*. Entende-se como fatores de *input* aqueles que são insumos para ocorrência do processo educativo. Scheerens (1990) cita como exemplos a experiência/formação do professor, a experiência do diretor e o suporte familiar.

Quanto aos fatores processuais, a literatura inicial das pesquisas de eficácia escolar sugeria um modelo de cinco fatores que combinados tornariam uma escola eficaz: *i*) liderança educacional forte; *ii*) ênfase na

aquisição das habilidades básicas; *iii*) ambiente organizado e seguro; *iv*) altas expectativas sobre o desenvolvimento dos estudantes; e *v*) avaliação frequente do progresso dos estudantes (ODDEN, 1982; SCHEERENS, 2000). Todavia, esses estudos iniciais foram muito criticados no que concerne ao método empregado.

Estudos posteriores, mais robustos metodologicamente, foram conduzidos, o que ampliou a lista de características das escolas eficazes. Sammons, Hillman e Mortimore (1995), na intenção de fornecer um resumo das evidências das pesquisas de eficácia escolar, identificaram onze fatores-chave. Quatro fatores relacionados à escola: ensino e objetivos claros, ambiente de aprendizagem ordenado e atraente, concentração no ensino e na aprendizagem (foco no desempenho) e organização orientada à aprendizagem (formação de pessoal na escola). Cinco fatores referem-se a características e ações dos coordenadores e professores: liderança profissional firme, objetiva e com enfoque participativo; objetivos e visões compartilhadas; altas expectativas e fornecimento de desafios; incentivo positivo e monitoramento constante do progresso. Há ainda um fator que envolve direitos e responsabilidades dos alunos (alta autoestima do aluno, oportunidade de assumir posições de responsabilidade, controle dos trabalhos) e um outro relacionado ao estabelecimento de relação de apoio e cooperação entre casa e escola.

Os autores ressaltaram, no entanto, que esses fatores não devem ser considerados de modo independente. Várias associações entre eles podem ajudar a fornecer um melhor entendimento de prováveis mecanismos de eficácia. Ressalta-se ainda que esses fatores foram extraídos de pesquisas de etapas de ensino distintas, considerando contextos e desempenhos específicos.

Quanto ao *output*, a maioria das pesquisas utiliza-se exclusivamente de resultados cognitivos dos alunos na área de leitura e matemática (SAMMONS; HILLMAN; MORTIMORE, 1995). O *output* deve ser o resultado esperado da escola: alunos bem formados. Certamente, a formação adequada do aluno implica mais do que a formação acadêmica cognitiva. Todavia, a ausência de medidas sistemáticas de outros aspectos que complementam a formação escolar faz com que as pesquisas utilizem majoritariamente como *output* o desempenho cognitivo.

Em consonância com os avanços teóricos, a análise das produções na área de eficácia escolar demonstra um sofisticado avanço metodológico. Inicialmente, alguns resultados controversos encontrados na literatura decorreram de dificuldades metodológicas que inflaram determinados resultados, uma vez que não utilizavam análises que consideram a estrutura hierárquica do sistema educacional. Por exemplo, procedimentos estatísticos tradicionais, tal como a análise de regressão múltipla, tratam os indivíduos de modo independente, não considerando a sua estrutura grupal e ignorando as possíveis influências dessa estrutura (GOLDSTEIN, 2010; HOX, 2010).

Outra questão que tem sido cada vez melhor tratada é a dificuldade de se distinguir os componentes estáveis e instáveis, uma vez que a maior parte dos estudos de eficácia escolar é transversal. A fim de resolver essa questão, além do aumento de estudos longitudinais, técnicas de análises de medidas repetidas têm ganhado espaço nas pesquisas (WILLMS; RAUDENBUSH, 1989). Pode-se citar, por exemplo, os estudos longitudinais desenvolvidos por Ferrão e Couto (2014) e Dumay, Coe e Anumendem (2014) que buscam avaliar valor agregado e estabilidade, respectivamente. Percebe-se uma sofisticação metodológica na área, sendo que cada vez mais têm sido utilizadas: modelagem multinível, meta-análises, modelagem por equações estruturais, modelagem por curva de crescimento, além de pesquisas que empregam multimétodos.

De modo a sintetizar a história de desenvolvimento da área de eficácia escolar, Reynolds et al. (2011) a divide em cinco fases. A primeira fase é marcada por um conjunto de estudos que buscam demonstrar que as escolas fazem a diferença, em oposição ao chamado relatório Coleman. Já a segunda e a terceira fases (meados da década de 1980 e início da década de 1990) são marcadas por avanços metodológicos na busca de melhor estimar o efeito-escola e pelas inúmeras tentativas de explorar as razões que explicam os diferentes efeitos entre escolas, respectivamente.

A quarta fase (final da década de 1990) é o período de internacionalização do campo de pesquisa e integração de resultados. É neste período que se iniciam as primeiras pesquisas no Brasil. As primeiras iniciativas brasileiras estão relacionadas ao início das avaliações em larga escala e os primeiros estudos foram publicados no início dos anos 2000.

A quinta e última fase (início dos anos 2000 e ainda corrente) inicia-se com os estudos que tentam analisar a eficácia escolar de forma dinâmica, compreendendo a educação como um conjunto de fatores que se relacionam. O modelo dinâmico busca explicar porque os sistemas educacionais funcionam diferenciadamente, baseando-se em quatro pressupostos: *i*) tempo e oportunidade são variáveis do nível do aluno que estão diretamente relacionadas ao sucesso; *ii*) ensino de qualidade e currículo são variáveis de processo que influenciam o tempo e a oportunidade de aprendizado; *iii*) por sua vez, ensino de qualidade, tempo e oportunidade são variáveis que sofrem influência de variáveis do nível da escola que podem contribuir para a promoção desses fatores nas salas de aula; e, por fim, *iv*) o sucesso no desempenho também é determinado por fatores atitudinais, motivacionais e comportamentais relacionados ao estudante. Além disso, quatro aspectos operam na relação entre essas variáveis: consistência, coesão, constância e controle (KYRIAKIDES, 2008).

Embora os estudos sobre eficácia escolar tenham se aprimorado em termos de fundamentação teórica, ainda se fazem necessários estudos que sustentem tais fundamentações, o que demanda também aprimoramentos metodológicos. No entanto, para avançar é essencial consolidar os resultados encontrados até o momento, mesmo porque uma das críticas à área é que, devido ao seu crescimento rápido, ela aprendeu pouco com ela mesma (REYNOLDS et al., 2011). Nesse sentido, este estudo tem como objetivo revisar criticamente as produções brasileiras na área de eficácia escolar. Especificamente, busca-se analisar as concepções teóricas, os objetivos, as variáveis (de contexto, *input*, processo e *output*), tipos de delineamento, modelos de análise e os principais resultados encontrados.

Decidiu-se por realizar uma revisão da literatura brasileira para que seja possível confrontar os resultados nacionais com os destacados na literatura estrangeira, sobretudo porque os fatores associados ao desempenho escolar podem sofrer influência das características do sistema educacional sob análise e da realidade de determinado país. Além disso, a discussão dos resultados brasileiros à luz da literatura estrangeira permitirá delinear e propor pesquisas futuras para essa área, que ainda necessita de subsídios acadêmicos para crescer.

Método

A fim de selecionar as produções sobre eficácia escolar no Brasil, a pesquisa bibliográfica foi realizada com as etapas e critérios descritos a seguir.

1) Busca na base de dados da Scielo. Essa busca foi realizada com a utilização das palavras-chave: *eficácia escolar* (5 artigos), *efeito escola* (2 artigos), *efeito-escola* com a utilização de hífen (3 artigos) e *desempenho escolar* (43 artigos). Dessa primeira busca, resultaram 52 artigos, pois houve a incidência de um artigo comum entre as palavras-chave *eficácia escolar* e *desempenho escolar*.

2) Leitura e classificação dos artigos. Para a inclusão do artigo nesta revisão de literatura, necessariamente este precisava abordar questões referentes a: estimação de efeito-escola ou valor agregado; eficácia ou eficiência escolar; equidade e igualdade; identificação de fatores escolares que contribuem para a eficácia escolar ou métodos adequados para avaliação da eficácia escolar. Esse critério foi estabelecido tendo como base a definição proposta por Ferrão e Couto:

> Eficácia escolar designa a área de investigação científica em educação que é dedicada à estimação do efeito-escola (nas dimensões de valor agregado, eficácia diferencial e equidade social), à identificação de fatores que contribuem para que uma escola seja eficaz e abrange ainda o campo de trabalho cujo enfoque é a procura de métodos adequados e fiáveis para medir qualidade da escola (2013, p. 137) .

Foram excluídos 36 artigos por não atenderem a esse critério.

3) Avaliação das referências bibliográficas dos 16 artigos classificados como adequados. Essa avaliação resultou na seleção de mais 11 artigos, dos quais dois foram excluídos por não se tratarem de relato de pesquisa ou de discussão sobre a literatura da área.

4) Avaliação das referências bibliográficas dos nove novos artigos agregados à seleção. Essa nova avaliação resultou na

identificação de mais cinco artigos que se enquadravam nos requisitos desta pesquisa.

5) Avaliação das referências bibliográficas dos cinco novos artigos agregados à seleção. Nessa avaliação não foram encontradas novas referências. Encerrou-se, assim, a pesquisa bibliográfica. Portanto, a busca iniciou-se na base da Scielo, mas foram integrados novos artigos até não serem encontrados outros novos, compreendendo que as principais produções da área seriam em algum momento citadas.

Para a análise dos artigos selecionados, optou-se por um procedimento padronizado para extração e classificação das informações referentes ao tipo de pesquisa (estudo teórico ou relato de pesquisa), à finalidade, à origem dos dados, ao tipo de análise e às variáveis utilizadas.

Resultados e Discussão

Foram levantados 30 artigos publicados em periódicos científicos entre 2000 e 2013. O ano em que ocorreu o maior número de publicação foi 2007 (n. = 6). Apesar de não ser uma produção densa, houve a publicação de pelo menos um artigo por ano ao longo dos últimos 14 anos. Nota-se que a primeira publicação brasileira ocorreu muito tempo depois das primeiras publicações na área de eficácia escolar. Esse fato está relacionado ao estabelecimento das avaliações externas no Brasil, importante fonte de dados para os estudos de eficácia escolar. A escala do Sistema Nacional de Avaliação da Educação Básica (Saeb), o primeiro e mais importante processo de monitoramento da qualidade da educação brasileira, se consolida em 1997 e, considerando os prazos entre aplicação, divulgação de microdados e realização da pesquisa, entende-se as primeiras publicações só virem a ocorrer no início dos anos 2000.

Quanto ao tipo de pesquisa, cinco artigos dentre os 30 selecionados são teóricos, o que representa 16% dos artigos selecionados no levantamento bibliográfico (FERRÃO; FERNANDES, 2003; FRANCO; BONAMINO, 2005; SOARES, 2007; FERRÃO, 2012; KOSLINSKI; ALVES, 2012). Os estudos teóricos apresentaram reflexões sobre aspectos da operacionalização de estudos

sobre eficácia escolar, sobre os modelos de análise e os caminhos para apropriação dos resultados e para compreensão das desigualdades de oportunidades. De modo geral, esses estudos teóricos tinham caráter mais sintetizador do que propositivo. Percebe-se uma carência na sistematização e na reflexão dos achados na área, bem como de desenvolvimento conceitual e teórico dos construtos investigados. Ressalta-se que uma das críticas à área de eficácia escolar é a existência de numerosos estudos empíricos, carentes de uma teoria sólida a ser verificada (VAN DEN EEDEN; HOX; HAUER, 1990).

Os 25 artigos restantes tratam de relatos de pesquisas. As informações gerais extraídas desses estudos empíricos estão apresentadas no quadro 1. Nota-se que todas as pesquisas foram realizadas na educação básica, sobretudo com a base de dados do Saeb, e utiliza-se de regressão multinível para análise dos dados. Essas duas características são condizentes com as pesquisas estrangeiras. A escolha pela educação básica é mais comum devido à existência de uma quantidade maior de informações, provindas de diversos sistemas avaliativos em larga escala, o que não é tão comum para a educação superior. Quanto ao uso de regressão multinível, essa é a técnica mais adequada, considerando a organização do sistema educacional em que alunos são agregados por escolas (HOX, 2010; LEE, 2008).

Quadro 1 Autores, ano de publicação, base de dados, etapa de ensino e tipo de análise dos relatos de pesquisa sobre eficácia escolar no Brasil

Autores	Ano	Base de dados	Série	Análise
Barbosa, Fernandes	2000	Saeb 1997	8ª EF[1]	RM[2]
Soares, Alves, Oliveira	2001	UFMG[3] 1998-2000	3º EM[4]	RM
Ferrão et al.	2001	Saeb 1999	4ª EF	RM
Albanez, Ferreira, Franco	2002	Saeb 1999	8ª EF	RM
Soares, Alves	2003	Saeb 2001	8ª EF	RM
Andrade, Franco, Carvalho	2003	Saeb 1999	3º EM	RM
Soares, T. M.	2003	Simave[5] 2002	4ª EF	RM
Soares (a)	2004	Saeb 2001	8ª EF	RM

Autores	Ano	Base de dados	Série	Análise
Soares (b)	2004	Saeb 2001	8ª EF	RM
Jesus, Laros	2004	Saeb 2001	8ª EF	RM
Soares, T. M.	2005	Simave 2002	4ª EF	RM
Soares, Andrade	2006	Simave 2002/ UFMG 2002-2004	3º EM	RM
Franco et al.	2007	Saeb 2001	4ª EF	RM
Andrade, Laros	2007	Saeb 2001	3º EM	RM
Nascimento	2007	Avaliação Estado da Bahia	4ª e 8ª EF	RL[6]
Alves, Soares (a)	2007	Coleta própria	5ª e 6ª EF	RM
Alves, Soares (b)	2007	Coleta própria	5ª EF	RM
Alves, Soares	2008	Coleta própria	5ª e 6ª EF	RM
Teixeira	2009	Projeto Geres[7]	Alfabetização	AQ[8]
Bonamino et al.	2010	PISA[9] 2000	15 anos	RL
Stocco, Almeida	2011	Projetos Geres e Nepo[10]	1ª – 4ª EF	AD[11]
Rodrigues et al.	2011	Saeb 1997-2005	4ª EF	RL/AC[12]
Silva, Bonamino, Ribeiro	2012	Coleta própria	EJA[13]	AQ
Laros, Marciano, Andrade	2012	Saeb 2001	3º EM	RM
Ferrão, Couto	2013	Projeto Geres	1ª – 4ª EF	RM

Notas:

1 EF – Ensino fundamental
2 RM – Regressão multinível
3 UFMG – Universidade Federal de Minas Gerais
4 EM – Ensino médio
5 Simave – Sistema Mineiro de Avaliação da Educação Pública
6 RL – Regressão linear
7 Geres – Estudo Longitudinal da Geração Escolar
8 AQ – Análise qualitativa (observações e entrevistas)
9 Pisa – Programme for International Student Assessment (Programa Internacional de Avaliação de Estudantes)
10 Nepo – Núcleo de Estudos de População da Unicamp
11 AD – Análise descritiva
12 AC – Análise contrafactual
13 EJA – Educação de Jovens e Adultos

Na análise multinível, a dependência entre as observações pode ser avaliada pela correlação intraclasse (ICC), que representa a homogeneidade em um mesmo grupo e ao mesmo tempo a heterogeneidade entre grupos distintos (LAROS; MARCIANO, 2008). Quando a ICC é superior a 10% da variância total, recomenda-se a utilização de um modelo hierárquico (LEE, 2008). Nas pesquisas levantadas, observa-se ICC entre 21 e 46% e, depois de corrigidas pelo contexto socioeconômico e cultural, entre 10 e 27%.

No caso do Brasil, controlar as características socioeconômicas, sociais e culturais nos estudos educacionais faz-se essencial diante das desigualdades existentes no país e da seletividade de composição das escolas. Como mencionado por Fletcher (1998, p. 3), "a composição da escola é fortemente condicionada pelas circunstâncias de seu contexto, especialmente a segregação residencial em seu entorno, provocada pelas diferenças socioeconômicas". Esse controle é ainda mais necessário diante dos resultados que demonstram o grande impacto que o nível socioeconômico exerce sobre o desempenho escolar. Caso contrário, seria superestimado o efeito daquelas escolas que recebem alunos de maior poder aquisitivo e social.

Esses resultados estão em consonância com o estudo de Ferrão e Fernandes (2003) que indicou uma ICC geral em torno de 30% e em torno de 19%, após controle contextual, no Brasil. Esses valores são bem superiores aos observados na maioria dos países desenvolvidos, que giram em torno de 19% e, quando ajustado para o contexto, em torno de 10% (RUTTER; MAUGHAN, 2002; ALVES; SOARES, 2007b). Contudo, estão bastante próximos dos valores encontrados nos estudos mexicanos, que variam entre 12% e 43% (CARVALLO-PONTÓN, 2010).

Dos 25 estudos empíricos, destaca-se que cinco possuem delineamento longitudinal: três deles utilizaram uma base de dados composta por sete escolas de Belo Horizonte (ALVES; SOARES, 2007a; ALVES; SOARES, 2007b; ALVES; SOARES, 2008) e dois utilizaram a base de dados do Projeto Geres (FERRÃO; COUTO, 2013; STOCCO; ALMEIDA, 2011). Esse projeto surgiu da parceria entre seis centros universitários com tradição em avaliação da educação, a saber: PUC-Rio, UFMG, UNICAMP, UFBA, UFJF e UEMS. Trata-se de um estudo longitudinal de painel que teve início em 2005, no qual a mesma amostra de escolas e de alunos dos anos iniciais foi observada ao longo de quatro anos.

O delineamento longitudinal é apontado como o mais adequado para a avaliação da eficácia escolar, sobretudo quando se trata de efeito-escola, uma vez que em estudos transversais a medida de desempenho representa o agregado de todo o aprendizado adquirido pelo estudante ao longo dos anos escolares e não o valor agregado pela escola cujos fatores associados estão sendo avaliados (FRANCO; BROOKE; ALVES, 2008). Dessa forma, surge um descompasso entre a medida de *output* e a medida das condições escolares. Entretanto, devido às dificuldades operacionais e políticas – tais como tempo, recursos financeiros, firmar e manter parcerias com instituições e manter o acompanhamento da amostra – os estudos longitudinais dentro e fora do Brasil ainda são em menor número.

Ao avaliar o objetivo dos 25 artigos empíricos brasileiros na área de eficácia escolar, é possível categorizá-los em três segmentos quanto ao seu objetivo: *i*) artigos que buscavam estimar o efeito-escola ou o valor agregado pela escola; *ii*) artigos que buscavam avaliar o quanto a escola tem contribuído para a equidade e igualdade social; e *iii*) artigos que buscavam investigar os fatores contextuais e escolares relacionados à eficácia escolar.

Artigos relacionados ao efeito-escola

Nesse primeiro segmento, encontram-se os cinco artigos de delineamento longitudinal e o artigo de Soares, Alves e Oliveira (2001). Neste último caso, para a estimação do efeito-escola, foi necessário adaptar uma medida indireta da habilidade acadêmica prévia a partir de informações do questionário acerca do histórico escolar. Essa adaptação foi necessária porque os estudos de efeito-escola buscam estimar o quanto a escola acrescentou ao aprendizado do aluno.

Dentre os artigos sobre efeito-escola, o estudo de Stocco e Almeida (2011) diferencia-se por ser o único a utilizar análise descritiva. Com base em dados do Projeto Geres e do Projeto de Estudo de Vulnerabilidade Sociodemográfica (Nepo – Unicamp), os autores discutiram a distribuição espacial das escolas e o desempenho escolar. Não foi encontrada uma recorrência clara em relação à disposição espacial das escolas quando relacionadas ao seu desempenho.

Os demais artigos utilizaram modelos multiníveis para estimação do efeito-escola. Em todos eles, foi mencionada a importância de se realizar o controle da composição social da escola. Como já ressaltado, esse

controle é necessário para que não seja atribuída à escola um efeito que advém de características sociais, culturais e econômicas dos alunos (SOARES; ALVES; OLIVEIRA, 2001). Em geral, as variáveis utilizadas para o controle da composição social foram: nível socioeconômico do aluno, nível socioeconômico da escola (média do nível socioeconômico dos alunos), nível de escolaridade dos pais, sexo e etnia.

Em consonância com a literatura estrangeira, os resultados das pesquisas brasileiras também demonstram que a maior parte da variação entre resultados escolares pode ser explicada por fatores extraescolares. Contudo, mesmo após utilizar variáveis de controle da composição social, os estudos apontaram que a escola faz diferença no desempenho dos estudantes e efeitos diferenciados podem ser encontrados a depender da área de conhecimento (ALVES; SOARES, 2007b).

Destaca-se ainda que, entre os artigos que investigaram o efeito-escola, três apresentaram uma abordagem quantitativa e qualitativa (*mixed methods*). Nos estudos de Alves e Soares (2007a), Alves e Soares (2007b), Alves e Soares (2008), após a identificação do efeito das escolas, esses resultados foram interpretados, utilizando-se dados das entrevistas realizadas com os professores.

Artigos relacionados à equidade e igualdade

Nesse segundo segmento, encontram-se cinco artigos. Esses estudos corroboram a hipótese de que há no sistema educacional brasileiro desigualdades em termos de desempenho e de oportunidade de acesso a recursos. Vários estudos têm demonstrado, por exemplo, diferenças de desempenho por gênero, raça e nível socioeconômico. Compreender melhor essas diferenças é o foco dos artigos que tratam de igualdade e equidade. Ressalta-se, no entanto, que os artigos brasileiros focam principalmente em aspectos nos quais a escola tem pouca interferência direta.

A produção científica no Brasil que avalia equidade e igualdade é pequena e bastante diversificada. Há um artigo que avalia diferenças em relação ao nível socioeconômico (SOARES; ANDRADE, 2006), um que avalia diferença entre raças (SOARES; ALVES, 2003), um que avalia diferença entre gênero (ANDRADE; FRANCO; CARVALHO, 2003) e dois que avaliam o quanto características escolares podem contribuir para a promoção da equidade (SOARES, 2004a; SOARES, 2004b).

O artigo de Soares e Alves (2003) mostra que existe grande diferença de desempenho entre alunos brancos e negros e que parte dessas diferenças se deve ao efeito do nível socioeconômico. Quando verificados os efeitos moderadores de variáveis escolares, notou-se que melhor qualificação docente, melhores salários e melhores equipamentos têm efeito significativo no sentido oposto ao desejado. Dessa forma, a melhora nas condições docentes e escolares tende a potencializar o efeito de raça sobre desempenho, aumentando o hiato no desempenho escolar de alunos brancos e negros.

Em outro artigo, Soares e Andrade (2006) verificaram o quanto as diferentes dependências administrativas do sistema escolar (federal, estadual, municipal e particular) têm atuado de modo a diminuir o efeito do nível socioeconômico sobre o desempenho. Os autores não encontraram efeito para as escolas estaduais, municipais e particulares. Já as escolas federais tenderam a aumentar o efeito do nível socioeconômico sobre o desempenho. Os autores apontaram ainda a inexistência de escolas com alta eficácia escolar e alta equidade.

De modo geral, os artigos sobre equidade têm demonstrado que os fatores produtores de melhora no desempenho acadêmico geram efeitos de intensidades diferentes entre os estudantes. Especificamente, alunos mais favorecidos tendem a usufruir mais das melhorias das condições escolares (SOARES, 2004b). Esses resultados são perversos para o sistema educacional, pois demonstram um aumento das desigualdades na medida em que se melhora a eficácia escolar.

Artigos relacionados aos fatores associados

Nesse terceiro segmento estão 14 artigos. O foco desses estudos é identificar políticas e práticas escolares que podem explicar o alto desempenho educacional dos estudantes. Dois desses artigos utilizaram uma abordagem qualitativa com uso de técnicas observacionais e de entrevistas: Teixeira (2009) e Silva, Bonamino e Ribeiro (2012). Esse tipo de abordagem é interessante na área de eficácia escolar, pois permite compreender mais profundamente determinados fenômenos e, dessa forma, corrobora para uma consolidação teórica.

No estudo de Teixeira (2009), tendo como base os resultados de estudos quantitativos que apontavam a importância do ambiente de aprendizado,

do clima acadêmico e da infraestrutura, buscou-se avaliar os espaços destinados ao desenvolvimento de habilidades de leitura em seis escolas que apresentavam desempenhos diferenciados. Os resultados apontaram que os espaços da escola refletiam diferentes concepções de práticas pedagógicas, e que nas escolas com melhores desempenhos era atribuída maior importância aos espaços e aos recursos escolares. Esse resultado mostra a relevância da execução de estudos qualitativos e quantitativos. Demonstra ainda o quanto alguns fatores apresentados na literatura são catalisadores ou inibidores de ambientes propícios ao desenvolvimento e, por isso, não podem ser interpretados de forma isolada (FERRÃO, 2012).

Já no estudo de Silva, Bonamino e Ribeiro (2012), buscou-se analisar três escolas da rede municipal do Rio de Janeiro que integram o programa de educação de jovens e adultos e que têm apresentado bons resultados. Os autores concluem que muitos dos fatores associados ao desempenho encontrados nos estudos relativos à educação regular são também válidos para a educação de jovens e adultos.

Dentre os estudos quantitativos, a abordagem econométrica utilizada por Rodrigues, Rios-Neto e Pinto (2011) se diferencia das demais. Os autores buscaram avaliar o impacto do nível socioeconômico no desempenho escolar avaliado pelo Saeb entre 1997 e 2005. Para tanto, os autores utilizaram um método de decomposição contrafactual que permite isolar a contribuição da variação na composição e retorno do nível socioeconômico sobre a variação na média e distribuição do desempenho escolar. Os resultados confirmam os achados da literatura que apontam o forte efeito do nível socioeconômico sobre o desempenho. Além disso, os resultados indicaram que a queda no desempenho em matemática no Saeb pode ser explicada pela democratização das oportunidades de acesso à escola e consequente queda no nível socioeconômico geral.

Com relação aos demais 11 estudos quantitativos, eles utilizaram em suas análises regressão linear simples ou regressão multinível (BARBOSA; FERNANDES, 2000; FERRÃO et al., 2001; ALBANEZ; FERREIRA; FRANCO, 2002; SOARES, 2003; JESUS; LAROS, 2004; SOARES, 2005; ANDRADE; LAROS, 2007; FRANCO et al., 2007; NASCIMENTO, 2007; BONAMINO et al., 2010; LAROS; MARCIANO; ANDRADE, 2012). A fim de melhor sumarizar os achados, as variáveis investigadas em cada

estudo foram categorizadas com base no modelo integrado de Scheerens (1990): contexto, *input*, processo e *output*. Os quadros 2, 3 e 4 apresentam os resultados dessa sumarização. Nos quadros, foram utilizados os sinais ↑, ↓, —, para representar, respectivamente, uma relação positiva, negativa ou não significativa com o desempenho. A quantidade de vezes que os sinais aparecem representa a quantidade de artigos que utilizaram a variável/fator.

Quadro 2 Fatores/variáveis contextuais utilizados nos estudos brasileiros de eficácia escolar e sua relação com desempenho

FATORES/VARIÁVEIS DE CONTEXTO	1	2	3	4	5	6	7	8	9	10	11
Nível socioeconômico do aluno	↓	↓	↓	—	—	↑	↑	↑	↑	↑	
Nível socioeconômico da turma/ escola	↑	↑	↑	↑	↑	↑	↑	↑	↑		
Sexo (masculino)	↑	↑	↑	↑	↓	↓					
Etnia (não branco)	↓	↓	↓	↓	↓	↓	↓				
Etnia agregada (não branco)	↓	↓	—								
Escolaridade do pai ou da mãe	↑	↑	↑	—							
Diálogo familiar	↑										
Status ocupacional dos pais	↑										
Recursos educacionais familiares	↑										
Disponibilidade de recursos culturais (livros) em casa	↑	↑									
Disponibilidade de recursos culturais em casa (agregado*)	↑										
Grau de urbanização do município	—										
Rendimento médio do município	—										
Garantia do mínimo constitucional na educação	—										
Gastos efetivos em educação no município	—										
Relação n° de docentes/n° de alunos	—										
Rede (particular)	↑	↑	↑								

* O termo agregado especifica que a variável era incialmente do nível 1 e foi agregada para o nível 2.

Dentre as variáveis de contexto, confirma-se a relação consistente e positiva do nível socioeconômico da escola sobre o desempenho do aluno já encontrada em estudos estrangeiros (COLEMAN et al., 1966; FLETCHER, 1998). Quanto ao nível socioeconômico do aluno, a sua relação com o desempenho parece não ser tão clara. Ocorre que, quando o nível socioeconômico do aluno é inserido em conjunto com o nível socioeconômico da escola, o efeito da primeira variável tende a ser bastante pequeno e muitas vezes negativo, isso porque essas variáveis apresentam-se altamente correlacionadas.

Outra variável que aponta relação consistente com desempenho é raça. Os resultados demonstram que alunos negros apresentam, em média, desempenho inferior ao dos alunos brancos. Essa relação mostra a desigualdade presente no sistema educacional brasileiro. Ferrão et al. (2001) já indicava no início dos anos 2000 a necessidade e a importância de se procurar explicação para o efeito de raça sobre o desempenho. Todavia, pouco se avançou sobre a compreensão desse efeito.

Quanto à variável sexo, apesar de parecer que os resultados são controversos, eles não são. Os diferentes sentidos (efeitos positivos e negativos) estão relacionados com a disciplina avaliada. Os estudos apontam que, quando o desempenho avaliado é em matemática, os meninos têm resultados melhores (BARBOSA; FERNANDES, 2000; ALBANEZ; FERREIRA; FRANCO, 2002; ANDRADE; LAROS, 2007; FRANCO et al., 2007). Quando o desempenho avaliado é em língua portuguesa, as meninas têm resultados melhores (SOARES, 2003; SOARES, 2005).

Vale destacar que a escolaridade dos pais, apesar de aparecer com efeito positivo em somente três estudos, em geral tem apresentado uma relação forte e consistente, sobretudo quando se trata da escolaridade da mãe. Ocorre que em muitos estudos essa variável é agregada na composição do fator de nível socioeconômico e, por isso, ela não é incluída isoladamente no modelo. Para as demais variáveis, ainda é arriscado afirmar alguma tendência.

Ao avaliar o quadro 3, ressalta-se o efeito negativo do atraso escolar sobre o desempenho nos oito estudos em que essa variável foi avaliada. Esse resultado aponta que, quando o estudante é reprovado em algum momento, ele tende a ter um desempenho inferior aos demais estudantes no decorrer da trajetória escolar, mesmo tendo cursado algumas vezes uma determinada série.

Quadro 3 Fatores/variáveis de *input* utilizados nos estudos brasileiros de eficácia escolar e sua relação com desempenho

FATORES/VARIÁVEIS DE *INPUT*	1	2	3	4	5	6	7	8	9	10	11
Número de alunos na escola	—										
Aluno frequentou a pré-escola	↑										
Atraso escolar/defasagem idade-série	↓	↓	↓	↓	↓	↓	↓	↓			
Bolsa Escola	↓										
Aluno trabalha	↓	↓	↓								
Gosta da disciplina	↑	↑	↑								
Abandono escolar	↓										
Insuficiência de recursos financeiros	↓										
Turno	—										
Atraso escolar/defasagem idade-série (agregado)	↓	↓	↓	↓							
Gosta da disciplina (agregado)	—										
Aluno trabalha (agregado)	↓	↓	↓								
Estado de conservação dos equipamentos	↑	↑	↑	↑							
Infraestrutura escolar	↑	↑	↑	↑							
Formação do professor	↑	↑	↑								
Existência de pessoal e recursos	↑	—									
Idade média dos professores	↓										

Outra informação relevante apresentada no quadro 3 é a influência positiva que a infraestrutura e o bom estado de conservação dos equipamentos escolares exercem sobre o desempenho. Essa informação diverge dos resultados encontrados na literatura estrangeira (ALBANEZ; FERREIRA; FRANCO, 2002). Isso decorre da precariedade do sistema educacional brasileiro que ainda não garante uma estrutura mínima para o adequado funcionamento da escola. Em países desenvolvidos, a variância nos quesitos referentes a infraestrutura e equipamentos é tão pequena que não é significativa para explicar as diferenças no desempenho acadêmico (MURILLO; ROMAN, 2011).

Quanto às variáveis processuais (quadro 4), nota-se uma carência de informações que nos permita afirmar quais variáveis da escola podem realmente fazer a diferença no desenvolvimento acadêmico dos estudantes. Os resultados mais consistentes são relativos a passar/fazer dever de casa e ao bom clima escolar.

Quadro 4 Fatores/variáveis de processo utilizados nos estudos brasileiros de eficácia escolar e sua relação com desempenho

Fatores/variáveis de processo	1	2	3	4	5	6	7	8	9	10	11
Fazer dever de casa	↑	↑	↑	↑							
Uso do computador para fazer dever de casa	↓	↓									
Faltas do aluno	↓										
Faltas do aluno (agregado)	↓										
Faltas do professor	↓	↓									
Passar dever de casa	↑	↑	↑	↑	↑						
Bom clima escolar	↑	↑	↑	↑	—						
Cobrança/incentivo dos pais	↓	↓	↓								
Ênfase em resolução de problemas	↑										
Liderança e trabalhos colaborativos	↑	↑	—								
Interesse e comprometimento do professor	↑	↑	↑								
Taxa de estudo	↑										
Comparação com outros colegas	↑										
Sistema de ensino em ciclos	↓										
Cobrança/incentivo dos pais (agregado)	↑	↓									
Programas de recuperação de notas	—										
Professor exigente em sala de aula	↑	↑									
Expectativa do professor	↑										

Considerando os fatores processuais que contribuem para a constituição de escolas eficazes levantados na revisão de literatura de Sammons, Hillman e Mortimore (1995), as três variáveis correspondem aos fatores: *ambiente de aprendizagem e ensino e objetivos claros*. A diferença é que os fatores utilizados na literatura internacional são mais abrangentes. Por exemplo, o *fator ensino e objetivos claros* envolve aspectos como clareza de propósitos, aulas bem estruturadas, ensino adaptável, entre outros que estão além do simples passar dever de casa.

De modo geral, é possível encontrar certa correspondência entre as variáveis/fatores encontradas na literatura brasileira (quadro 4) e os 11 fatores sumarizados pela revisão de literatura de Sammons, Hillman e Mortimore (1995), por exemplo: bom clima escolar e liderança e trabalhos colaborativos. Por outro lado, outros fatores importantes não foram explorados como ensino e objetivos claros, monitoramento do progresso e organização orientada à aprendizagem, o que demonstra a necessidade de se investir na construção de medidas mais adequadas para avaliar as características processuais das escolas.

Por fim, em relação ao critério de *output* utilizado nos estudos quantitativos, três estudos utilizaram o desempenho em matemática, cinco em língua portuguesa, um realizou as análises para matemática e língua portuguesa em separado e dois consideraram no modelo o desempenho em diversas disciplinas. Destaca-se que todos os estudos quantitativos utilizaram como critério uma medida acadêmica. Para Soares e Alves (2003), isso não significa privilegiar o domínio cognitivo em relação aos demais domínios que devem ser desenvolvidos na escola, mas sim reconhecer a importância das competências cognitivas para se atingir outros objetivos, além desse domínio ser bastante dependente da estrutura escolar. Acrescenta-se ainda a ausência de medidas sistemáticas referentes aos outros domínios, o que dificulta a realização de pesquisas.

Conclusões e direções para pesquisas futuras

A vasta produção na área de eficácia escolar tem apontado para um conjunto de fatores (contextuais, organizacionais, de monitoramento e pedagógicos) que tornam a escola eficaz ao contribuir com o desenvolvimento do estudante, feita a ressalva de que não se pode esperar que ela elimine

completamente as inequidades sociais e biológicas (RUTTER; MAUGHAN, 2002). De modo geral, os resultados das pesquisas brasileiras convergem para os resultados observados em outros países. No entanto, fatores como infraestrutura e estado de conservação dos equipamentos ainda fazem diferença sobre o desempenho acadêmico no Brasil (FERRÃO et al., 2001; ALBANEZ; FERREIRA; FRANCO, 2002; JESUS; LAROS, 2004; ANDRADE; LAROS, 2007; FRANCO et al., 2007). Nesse contexto, na ausência de elementos tão básicos, torna-se difícil identificar efeitos homogêneos e significativos referentes a fatores processuais como práticas pedagógicas.

Para Murillo e Román (2011), esses resultados demonstram a necessidade de ajuste dos modelos de eficácia escolar para populações mais precárias ou países em desenvolvimento, para que isso não culmine na aplicação de políticas inadequadas destinadas a países em outro patamar de desenvolvimento educacional. Portanto, além de ainda permanecer em aberto a questão de como promover mudanças apropriadas naquelas escolas que não estão funcionando adequadamente, modelos teóricos específicos para países em desenvolvimento com problemas de equidade precisam ser desenvolvidos.

A Oficina Regional de Educação para a América Latina e o Caribe (OREALC/UNESCO) tem investido no desenvolvimento de ações voltadas para a produção de informação contextualizada sobre a aprendizagem dos alunos e a análise dos fatores associados ao desempenho, por exemplo, com as aplicações de três estudos comparativos (PERCE, SERCE e TERCE). Murillo (2008) destaca que Chile, México, Colômbia, Argentina e Brasil são os países com mais pesquisas na área de eficácia escolar e reforça que é preciso investir na sua continuidade.

Nota-se uma tendência em tornar a área mais teoricamente dirigida, com a proposição de modelos que integrem várias perspectivas de pesquisa (insumos, eficácia escolar, eficácia docente, equidade, igualdade, avaliação de sistemas) de modo a realizar uma análise mais global e de interação entre as variáveis (SCHEERENS, 2000). A análise dessa interação, em conjunto com a análise de aspectos como consistência, coesão, constância e controle, propostos, por exemplo, no modelo dinâmico, pode ser um dos caminhos para melhor compreender quais mudanças são mais efetivas (KYRIAKIDES, 2008).

É preciso mencionar que parte da dificuldade em se validar modelos e desenvolver uma fundamentação teórica da área está na construção das

medidas, sobretudo daquelas referentes aos aspectos processuais da escola. Faltam consistência e definições claras dessas medidas (KYRIAKIDES, 2008). Por exemplo, construtos bastante complexos como práticas pedagógicas muitas vezes se reduzem a uma variável "fazer dever de casa". Certamente, parte do problema de medida está igualmente associado à ausência de informações nos questionários das avaliações em larga escala, os quais devem ser aprimorados (KARINO; VINHA; LAROS, 2014).

A partir da revisão bibliográfica, fica claro também que, entre os desafios da área de eficácia escolar, está o delineamento e a execução de pesquisas longitudinais (REYNOLDS et al., 2011). A estimação do efeito-escola é mais adequada por meio de um delineamento em que os mesmos estudantes são acompanhados ao longo do tempo. Porém, diante da dificuldade de condução de estudos longitudinais, é preciso investir, em paralelo, em outras estratégias de análise que possibilitem a estimação tanto do efeito-escola quanto das variáveis que contribuem para a promoção da eficácia escolar, como o uso de outras medidas para estabelecimento de linha de base e o uso de medidas repetidas no nível da escola.

Felizmente, uma das tendências que parece surgir nos artigos sobre eficácia escolar é a realização de estudos de abordagem *quanti* e *quali* ou estudos multimétodo. Os estudos quantitativos são subsidiários de informações que permitem identificar sistemas modelo e fatores que parecem fazer a diferença. Tais resultados podem melhor orientar estudos com abordagem qualitativa, que permitem uma visão mais analítica de como as escolas funcionam e, consequentemente, podem contribuir para melhor compreensão do fenômeno. Assim, estimula-se tanto quanto possível a realização de estudos com abordagem multimétodo e a parceria de estudos *quanti* e *quali*.

Outro desafio da área é encontrar respostas para a promoção de escolas mais eficazes e mais equânimes. Os resultados estrangeiros e brasileiros têm consistentemente apontado que há desigualdade entre estudantes por nível socioeconômico, raça e gênero. Mais desanimador ainda são os resultados dos estudos sobre equidade que mostram que, ao se promover eficácia escolar, se promove um aumento das desigualdades, uma vez que estudantes mais favorecidos tendem a usufruir mais das melhorias no ambiente escolar. Debruçar-se sobre a problemática de como alcançar um sistema educacional mais igual e equânime é um desafio urgente.

Mesmo diante de grandes avanços ocorridos na área de avaliação educacional e nos estudos sobre eficácia escolar, há ainda muitas perguntas não respondidas. Buscou-se, a partir da análise das publicações brasileiras em artigos científicos e da confrontação com os resultados da literatura estrangeira, demonstrar necessidades, limitações e lacunas da área de eficácia escolar, bem como indicar caminhos para o crescimento e aprimoramento da produção científica. Espera-se, assim, que essa sumarização e reflexão sobre os achados instiguem pesquisadores e provoquem uma nova agenda de pesquisa.

Referências

ALBANEZ, A.; FERREIRA, F.; FRANCO, F. Qualidade e equidade no ensino fundamental brasileiro. *Pesquisa e Planejamento Econômico*, v. 32, n. 3, p. 453-475, dez. 2002.

ALVES, M. T. G.; SOARES, J. F. As pesquisas sobre os efeitos das escolas: contribuições metodológicas para a sociologia da educação. *Sociedade e Estado*, v. 22, n. 2, p. 435-473, jun. 2007a.

ALVES, M. T. G.; SOARES, J. F. Efeito-escola e estratificação escolar: o impacto da composição de turmas por nível de habilidade dos alunos. *Educação em Revista*, v. 45, p. 25-58, jun. 2007b.

ALVES, M. T. G.; SOARES, J. F. O efeito das escolas no aprendizado dos alunos: um estudo com dados longitudinais do Ensino Fundamental. *Educação e Pesquisa*, v. 34, n. 3, p. 527-544, set./dez. 2008.

ANDRADE, J. M.; LAROS, J. A. Fatores associados ao desempenho escolar: um estudo multinível com os dados do Saeb/2001. *Psicologia: Teoria e Pesquisa*, v. 23, n. 1, p. 33-42, jan./mar. 2007.

ANDRADE, M., FRANCO, C.; CARVALHO, J. B. P. Gênero e desempenho em matemática ao final do ensino médio: quais as relações? *Estudos em Avaliação Educacional*, n. 27, p. 77-95, jan./jun. 2003.

BARBOSA, M. E.; FERNANDES, C. Modelo multinível: uma aplicação a dados de avaliação educacional. *Estudos em Avaliação Educacional*, n. 22, p. 135-153, 2000.

BONAMINO, A.; ALVES, F.; FRANCO, C.; CAZELLI, S. Os efeitos das diferentes formas de capital no desempenho escolar: um estudo à luz de Bourdieu e de Coleman. *Revista Brasileira de Educação,* v. 15, n. 45, p. 487-594, set./dez. 2010.

CARVALLO-PONTÓN, M. Eficacia escolar: antecedentes, hallazgos y futuro. *Revista Internacional de Investigación en Educación,* v. 3, n. 5, p. 199-214, jul./dez. 2010.

COLEMAN, J. S.; CAMPBELL, E.; HOBSON, C.; MCPARTLAND, J.; MOOD, A.; WEINFELD, R.; YORK, R. *Equality of Educational Opportunity.* Washington, DC: US Department of Health, Education & Welfare, 1966.

DUMAY, X.; COE, R.; ANUMENDEM, D. N. Stability over time of different methods of estimating school performance. *School Effectiveness and School Improvement,* v. 25, n. 1, p. 64-82, 2014.

EDMONDS, R. Effective schools for the urban poor. *Educational Leadership,* v. 37, n. 1, p. 15-27, out. 1979.

FERRÃO, M. E. Avaliação educacional e modelos de valor acrescentado: tópicos de reflexão. *Educação & Sociedade,* v. 33, n. 119, p. 455-469, abr./jun. 2012.

FERRÃO, M. E.; COUTO, A. Indicador de valor acrescentado e tópicos sobre a consistência e estabilidade: uma aplicação ao Brasil. *Ensaio: Avaliação de Políticas Públicas Educacionais,* v. 21, n. 78, p. 131-164, jan./mar. 2013.

FERRÃO, M. E.; COUTO, A.P. The use of a value-added model for educational improvement: A case study from the portuguese primary education system. *School Effectiveness and School Improvement*, v. 25, n. 1, p. 174-190, 2014.

FERRÃO, M. E.; FERNANDES, C. O. O efeito-escola e a mudança: dá para mudar? Evidências da investigação brasileira. *Revista Electronica Iberoamericana sobre Calid, Eficaácia y Cambio en Educación,* v. 1, n. 1, p. 1-13, 2003.

FERRÃO, M. E.; BELTRÃO, K. I.; FERNANDES, C.; SANTOS, D.; SUAREZ, M.; ANDRADE, A. DO C. O Saeb – Sistema Nacional de Avaliação da Educação Básica: objetivos, características e contribuições na investigação da escola eficaz. *Revista Brasileira de Estudos de População,* v. 18, n. 1/2, p. 111-130, jan./dez. 2001.

FLETCHER, P. R. À procura do ensino eficaz. Relatório técnico. Brasília: PNUD/MEC/DAEB, 1998.

FRANCO, C.; BONAMINO, A. A pesquisa sobre característica de escolas eficazes no Brasil: breve revisão dos principais achados e alguns problemas em aberto. *Educação on-line: Revista do Programa de Pós-graduação em Educação,* v. 1, p. 1-13, 2005.

FRANCO, C.; BROOKE, N.; ALVES, F. Estudo longitudinal sobre qualidade e equidade no ensino fundamental brasileiro: GERES 2005. *Ensaio: Avaliação de Políticas Públicas Educacionais*, v. 16, n. 61, p. 625-638, out./dez. 2008.

FRANCO, C., ORTIGÃO, I., ALBERNAZ, A., BONAMINO, A., AGUIAR, G., ALVES, F., SÁTYRO, N. Qualidade e equidade em educação: reconsiderando o significado de "fatores intraescolares". *Ensaio: Avaliação de Políticas Públicas Educacionais*, v. 15, n. 55, p. 277-298, abr./jun. 2007.

GOLDSTEIN, H. *Multilevel Statistical Models* (4th edition). London: Wiley, 2010.

HOX, J. *Multilevel analysis: Techniques and applications*. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2010.

JESUS, G. R.; LAROS, J. A. Eficácia escolar: regressão multinível com dados de avaliação em larga escala. *Avaliação Psicológica*, v. 3, n. 2, p. 93-106, 2004.

KARINO, C. A.; VINHA, L. G. DO A.; LAROS, J. A. Os questionários do Saeb: O que eles realmente medem? *Estudos em Avaliação Educacional,* v. 25, n. 59, p. 270-297, set./dez. 2014.

KOSLINSKI, M. C.; ALVES, F. Novos Olhares para as desigualdades de oportunidades educacionais: A segregação residencial e a relação favela--asfalto no contexto carioca. *Educação & Sociedade*, v. 33, n. 120, p. 805-831, jul./set. 2012.

KYRIAKIDES, L. Testing the validity of the comprehensive model of educational effectiveness: A step towards the development of a dynamics model of effectiveness. *School effectiveness and school improvement,* v. 19, n. 4, p. 429-446, nov. 2008.

LAROS, J. A.; MARCIANO, J. L. Análise multinível aplicada aos dados do NELS 88. *Estudos em Avaliação Educacional*, v. 19, n. 40, p. 263-278, mai./ago. 2008.

LAROS, J. A.; MARCIANO, J. L.; ANDRADE, J. M. Fatores associados ao desempenho escolar em Português: um estudo multinível por regiões. Ensaio: Avaliação de Políticas Públicas Educacionais, v. 20, n. 77, p. 623-646, out./dez. 2012.

LEE, V. L. Utilização de modelos lineares hierárquicos lineares para estudar contextos sociais: o caso dos efeitos da escola. In: BROOKE, N.; SOARES, J. F. (eds.), Pesquisa em eficácia escolar: origem e trajetórias. Belo Horizonte: Editora UFMG, 2008, p. 273-296.

MORTIMORE, P. The nature and findings of school effectiveness research in primary sector. In: RIDDELL, S.; PECK, E. (Org.). School effectiveness research: its message for school improvement. Londes: HMSO, 1991.

MORTIMORE, P.; SAMMONS, P.; STOLL, L.; LEWIS, D.; ECOB, R. *School Matters: The Junior Years*. Shepton Mallett: Open Books, 1988.

MURILLO, F. J. *Enfoque, situación y desafios de la investigación sobre eficacia escolar em América Latina y el Caribe*. Eficacia escolar y factores associados en América Latina y el Caribe. Santiago, Chile: UNESCO, 2008, p. 7-47.

MURILLO, F. J.; ROMÁN, M. School infrastructure and resources do matter: analysis of the incidence of school resources on the performance of Latin American students. *School effectiveness and school improvement*, v.22, n.1, p. 29-50, fev. 2011.

NASCIMENTO, P. A. M. M. Desempenho escolar e gastos municipais por aluno em educação: relação observada em municípios baianos para o ano 2000. *Ensaio: Avaliação de Políticas Públicas Educacionais*, v. 15, n. 56, p. 393-412, jul./set. 2007.

ODDEN, A. Schools can improve: Local strategies need state backing. *State Education Leader,* p. 1-3, Summer, 1982.

REYNOLDS, D.; TEDDLIE, C.; CREEMERS, B.; SCHEERENS, J. TOWNSEND, T. An introduction to school effectiveness research. In: TEDDLIE, C.; REYNOLDS, D. (Org.), *The international handbook of school effectiveness research.* New York: Routledge, 2000, p. 3-25.

REYNOLDS, D.; SAMMONS, P.; FRAINE, B. D.; TOWNSEND, T.; DAMME, J. V. *Educational effectiveness research (EER): A state of the art review*. In: International Congress for School Effectiveness and Improvement, Cyprus, 2011.

RODRIGUES, C. G.; RIOS-NETO, E. L. G.; PINTO, C. C. DE X. Diferenças intertemporais na média e distribuição do desempeno escolar no Brasil: o papel do nível socioeconômico, 1997 a 2005. *Revista Brasileira de Estudos de População*, v. 28, n. 1, p. 5-36, jan./jun. 2011.

RUTTER, M.; MAUGHAN, B. School Effectiveness findings 1979-2002. *Journal of school psychology,* v. 40, n. 6, p. 451-475, 2002.

RUTTER, M.; MAUGHAN, B.; MORTIMORE, P.; OUSTON, J.; SMITH, A. Conclusões, especulações e implicações. In: BROOKE, N.; SOARES, J. F. (eds.). *Pesquisa em eficácia escolar:* origem e trajetórias. Belo Horizonte: Editora UFMG, 1979. p. 225-251.

SAMMONS, P.; HILLMAN, J.; MORTIMORE, P. *Key characteristics of effective schools: A review of school effectiveness research*. London: Office for Standards in Education [OFSTED], 1995.

SCHEERENS, J. School effectiveness research and the development of process indicators of school functioning. *School Effectiveness and School Improvement,* v. 1, n. 1, p. 61-80, 1990.

SCHEERENS, J. *Improving school effectiveness* (Fundamentals of Educational Planning N°. 68). Paris: UNESCO/International Institute for Educational Planning, 2000.

SILVA, J.; BONAMINO, A. M. C.; RIBEIRO, V. M. Escolas eficazes na educação de jovens e adultos: Estudo de casos na rede municipal do Rio de Janeiro. *Educação em Revista*, v. 28, n. 2, p. 367-392, jun. 2012.

SOARES, J. F. Qualidade e equidade na educação básica brasileira: a evidência do Saeb-2001. *Archivos Analíticos de Políticas Educativas*, v. 12, n. 38, p. 1-28, ago. 2004a.

SOARES, J. F. O efeito da escola no desempenho cognitivo de seus alunos. *Revista Electronica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación,* v. 2, n. 2, p. 83-104, 2004b.

SOARES, J. F. Melhoria do desempenho cognitivo dos alunos no ensino fundamental. *Cadernos de Pesquisa,* v. 37, n. 130, p. 135-160, jan./abr. 2007.

SOARES, J. F.; ALVES, M. T. G. Desigualdades raciais no sistema brasileiro de educação básica. *Educação e Pesquisa,* v. 29, n. 1, p. 147-165, jan./jun. 2003.

SOARES, J. F.; ANDRADE, R. J. Nível socioeconômico, qualidade e equidade das escolas de Belo Horizonte. *Ensaio: Avaliação de Políticas Públicas Educacionais,* v. 14, n. 50, p. 107-126, jan./mar. 2006.

SOARES, J. F.; ALVES, M. T. G.; OLIVEIRA, R. M. O efeito de 248 escolas de nível médio no vestibular da UFMG nos anos de 1998, 1999 e 2000. *Estudos em Avaliação Educacional*, n. 24, p. 69-117, jul./dez. 2001.

SOARES, T. M. Influência do professor e do ambiente em sala de aula sobre a proficiência alcançada pelos alunos avaliados no SIMAVE 2002. *Estudos em Avaliação Educacional*, n. 28, p. 103-124, jul./dez. 2003.

SOARES, T. M. Modelo de 3 níveis hierárquicos para a proficiência dos alunos de 4 serie avaliados no teste de língua portuguesa do SIMAVE/PROEB 2002. *Revista Brasileira de Educação,* v. 29, p. 73-88, mai./ago. 2005.

STOCCO, S.; ALMEIDA, L. C. Escolas municipais de Campinas e vulnerabilidade sociodemográfica: primeiras aproximações. *Revista Brasileira de Educação*, v. 16, n. 48, p. 663-814, set./dez. 2011.

TEIXEIRA, R. A. Espaços, recursos escolares e habilidades de leitura de estudantes da rede pública municipal do Rio de Janeiro: estudo exploratório. *Revista Brasileira de Educação,* v. 14, n. 41, p. 232-390, mai./ago. 2009.

TEODOROVIC, J. School Effectiveness: Literature review. Зборник Института за педагошка истраживања (*Serbian Institute for Educational Research*), v. 41, n. 1, p. 297-314, jul. 2009.

VAN DEN EEDEN, P.; HOX, J.; HAUER, J. Theory and model in multilevel research: Convergence or divergence? Amsterdam: SISWO, 1990.

WILLMS, J. D. *Monitoring school performance*. Washington, D.C.: The Falmer Press, 1992.

WILLMS, J. D.; RAUDENBUSH, S. W. A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement*, v. 26, n. 3, p. 209-232, Fall, 1989.

Camila Akemi Karino
Doutora em Psicologia pela Universidade de Brasília
Diretora de Avaliação do Geekie, Brasil
camilaakarino@gmail.com

Jacob Arie Laros
Ph.D. em Psicologia pela University of Groningen, Holanda
Professor da Universidade de Brasília
jalaros@gmail.com

A AVALIAÇÃO NO PLANO NACIONAL DE EDUCAÇÃO – PNE (2014 – 2024)

THE EVALUATION IN THE NATIONAL EDUCATION PLAN – PNE (2014 – 2024)

LA EVALUACIÓN EN EL PLAN NACIONAL DE EDUCACIÓN – PNE (2014-2024)

Catarina de Almeida Santos

Danielle Xabregas Pamplona Nogueira

RESUMO

Este artigo analisa o tema da avaliação no Plano Nacional de Educação – PNE (2014-2024), buscando compreender a concepção de avaliação presente no plano e sua estreita relação com a concepção de educação apresentada além de trazer alguns apontamentos quanto ao Sistema Nacional de Avaliação da Educação Básica (Sinaeb). Concluiu-se que a concepção de educação no PNE se mantém vinculada ao dispositivo constitucional da educação de qualidade como direito de todos. A avaliação no PNE se assume de forma mais sistêmica e mais ampla, servindo como subsídio ao plano e ao desenvolvimento de políticas educacionais apesar da vinculação aos resultados do Índice de Desenvolvimento da Educação Básica (Ideb). Em relação ao Sinaeb, verificaram-se avanços quanto à concepção de avaliação existente, embora tenham sido mantidas as avaliações já desenvolvidas no Brasil. Por fim, destacaram-se as incertezas da implementação das diretrizes do PNE quanto à avaliação, decorrentes da descontinuidade política brasileira, e dois modelos de avaliação presentes nas políticas avaliativas.

Palavras-chave: Plano Nacional de Educação; avaliação; Sinaeb.

ABSTRACT

This article analyzes the theme of evaluation in the National Education Plan – NEP (2014-2024), in order to understand the conception of evaluation present in the plan and its relationship with the conception of education, and to show some notes about the National System of Basic Education Evaluation (Sinaeb). It was concluded that the conception of education in the PNE reinforces the constitutional provision of quality education as a right of all. Evaluation in the PNE takes a more systemic

and broader form, serving as subsidy to the plan and to the development of educational policies, despite the linkage to Ideb's results. As for Sinaeb, progress has been made, although it maintains the evaluation programs that have already been developed in Brazil. Finally, the uncertainties of the PNE implementation have been highlighted, due to the Brazilian political discontinuity, and two evaluation models were presented in the evaluation policies.

Keywords: National Education Plan; evaluation; Sinaeb.

RESUMEN

Este artículo examina el tema de la evaluación dentro del Plan Nacional de Educación – PNE (2014-2024), tratando de entender el diseño de esta evaluación dentro del plan y su estrecha relación con el diseño de la educación anunciada, además de plantear algunas observaciones sobre el sistema nacional de evaluación de la educación básica – Sinaeb. Se concluyó que el diseño de la educación en el PNE se mantiene vinculada al dispositivo constitucional de la educación de calidad como un derecho para todos. La evaluación en el PNE se asume de forma más sistémica y más amplia, y sirve como subsidio para planear y desarrollar las políticas educativas, a pesar de la vinculación a los resultados del Ideb. En cuanto al Sinaeb, ha habido avances en relación con el diseño de la evaluación existente, aunque mantenga las evaluaciones ya desarrolladas en Brasil. Por último, se destacó la incertidumbre de la implementación de las directrices del PNE en relación a la evaluación, mediante la discontinuidad política brasileña y dos modelos de evaluación presentes en la política de evaluación.

Palabras clave: Plan Nacional de Educación; evaluación; Sinaeb.

Introdução

A Constituição Federal de 1988 (CF/1988), em tempos de redemocratização do Estado brasileiro, retomou em seus dispositivos a educação como direito de todos. Foi definido como princípio de ensino a garantia do padrão de qualidade educacional. Para orientar quanto a diretrizes, objetivos, metas e estratégias de implementação que assegurem a manutenção e o desenvolvimento do ensino em seus diversos níveis, etapas e modalidades, e que conduzam à melhoria da qualidade do ensino, entre

outros, a constituição determinou a elaboração do Plano Nacional de Educação (PNE), de duração decenal.

O PNE (2014 –2024), sancionado pela Lei nº 13.005, de 25 de junho de 2014, teve seu lócus privilegiado de articulação e debate na Conferência Nacional de Educação (Conae), que aconteceu em março de 2010. Nessa conferência, os debates versaram, especialmente, acerca da concepção de educação que deveria dar o tom de políticas, programas e ações educacionais na próxima década.[1]

Nesse sentido, falar de avaliação no PNE implica trazer à baila o princípio norteador do plano, o qual permeia suas metas: a educação de qualidade e as condições para concretização dessa qualidade. Assim, o presente texto tem como objetivo fazer uma análise a respeito do tema da avaliação no PNE, buscando compreender qual a concepção de avaliação presente no plano e sua estreita relação com a concepção de educação anunciada. Objetiva, ainda, trazer alguns apontamentos quanto ao Sistema Nacional de Avaliação da Educação Básica (Saeb), conforme disposto no PNE. Para o alcance dos objetivos propostos, utilizou-se, como perspectiva metodológica, a análise documental, com base em duas fontes principais: o documento da Conae 2010 e o PNE 2014-2024.

Qualidade da educação e avaliação no Documento Referência da Conae (2010)

A Conae de 2010 elegeu a qualidade da educação não só como um dos seus eixos, mas como princípio norteador do Documento Referência, o qual serviria para subsidiar a elaboração do próximo PNE, já que o último teve duração de 2001 a 2010. A conferência foi referendada pela compreensão de que a educação é um direito social, conforme dispositivo constitucional, e que a garantia desse direito só se efetiva por meio de sua oferta com qualidade a todos, independentemente de condições sociais e econômicas, raça, cor, orientação sexual, gênero ou etnia.

1 Embora o texto encaminhado pelo executivo não tenha traduzido as deliberações da Conae 2010, tais deliberações foram fundamentais no processo de debate, alteração e aprovação do documento final do PNE.

Em uma perspectiva ampliada, a concepção de educação que norteou o documento foi a de que ela é elemento partícipe das relações sociais e pode, assim, contribuir para a transformação ou para a manutenção dessas relações. A educação é uma prática social e cultural que tem como lócus privilegiado, mas não exclusivo, instituições educativas, espaços de difusão, criação e recriação cultural, e espaços de investigação sobre a efetivação do processo educativo experimentado por educandos e educadores na garantia de seus direitos.

No que se refere à qualidade da educação, sua compreensão deve se dar em uma perspectiva polissêmica, pois trata-se de um conceito complexo e que requer parâmetros comparativos para o que se julga bom ou ruim nos fenômenos sociais e educativos. É preciso compreender, ainda, que a qualidade e seus parâmetros compõem o sistema de valores da sociedade, que variam de acordo com cada momento histórico e com circunstâncias temporais e espaciais. Dessa forma, por se tratar de uma construção humana, a concepção de qualidade está diretamente vinculada ao projeto de sociedade, às relações sociais e à correlação de forças; é, portanto, produto dos confrontos e dos acordos de grupos e classes, que dão concretude ao tecido social em cada realidade.

Assim, os embates travados no âmbito da Conae se desenrolaram no sentido de garantir uma educação emancipadora, e, como tal, pactuou-se que o sentido de qualidade, para essa educação, decorre do desenvolvimento das relações sociais (políticas, econômicas, históricas, culturais), de modo que os homens sejam sujeitos de suas ações e os processos sejam definidos por eles de forma participativa e sustentável. O citado documento faz uma defesa inconteste de que os processos educacionais de crianças, jovens, adultos e idosos têm que contribuir para a apropriação das condições de produção cultural e de conhecimentos, e sua gestão, para o fortalecimento da educação pública e privada, a fim de construir uma relação efetivamente democrática.

Nessa ótica, a qualidade da educação que se quer ter envolve a formação para a emancipação dos sujeitos sociais, sem guardar, em si, critérios delimitantes. A educação de qualidade é, nessa perspectiva, aquela que contribui com a formação do estudante em seus aspectos humanos,

estéticos, sociais, culturais, filosóficos, científicos, históricos, antropológicos, afetivos, econômicos, ambientais e políticos, para que ele possa desempenhar seu papel de homem e cidadão no mundo. Essa é, assim, uma qualidade referenciada no social.

Dourado, Oliveira e Santos (2007), ao discutirem sobre os diversos atores implicados na garantia da educação de qualidade, assim como sobre as condições necessárias para a efetivação desta, apontam que ela envolve condições intra e extraescolares, bem como atores individuais e institucionais. Segundo eles:

> A discussão sobre Qualidade da Educação implica o mapeamento dos diversos elementos para qualificar, avaliar e precisar a natureza, as propriedades e os atributos desejáveis ao processo educativo, tendo em vista a produção, organização, gestão e disseminação de saberes e conhecimentos fundamentais ao exercício da cidadania e, sobretudo, a melhoria do processo ensino-aprendizagem (DOURADO; OLIVEIRA; SANTOS, 2007, p. 24).

Diante dos conceitos de educação e qualidade que permeiam o documento da Conae, buscou-se refletir sobre a concepção de avaliação que qualifica a educação nessa perspectiva.[2]

O termo "avaliação" é derivado das palavras "valor" e "ação"; o que implica dizer, portanto, que traz uma concepção valorativa, nesse caso, da ação educacional. Segundo Casali (2007, p. 10), avaliação é, "de modo geral, como saber situar cotidianamente, numa certa ordem hierárquica, o valor de algo enquanto meio (mediação) para a realização da vida do(s) sujeitos(s) em questão, no contexto dos valores culturais e, no limite, dos valores universais".

No documento da Conae (2010), "situam-se a avaliação da educação e a necessária articulação entre a concepção de avaliação formativa, indicadores de qualidade e a efetivação de um subsistema nacional de avaliação da educação básica e superior" (p. 52). Desse modo, o documento reforça os dispositivos legais, como a CF/1988 e a Lei de Diretrizes e Bases (Lei nº

2 Evidentemente que há aspectos da qualidade da educação que não são mensuráveis nos instrumentos de medidas educacionais, mas que fazem parte dos processos educativos e devem ser levados em conta no planejamento das políticas e das ações em nível de sistema, de escola e de sala de aula.

9.394/1996), como elementos para a garantia da qualidade da educação, e a avaliação como base para a melhoria dos processos educativos. Além disso, aponta a necessidade de se definir "competências dos entes federativos, especialmente da União, visando assegurar o processo nacional de avaliação das instituições de educação, com a cooperação dos sistemas de ensino" (BRASIL, 2010, p. 54). Isso porque, segundo o documento:

> Ao adotar a avaliação como eixo de suas políticas, o Brasil não o faz por meio de um sistema nacional, que envolva a educação básica e superior, mas desenvolve ações direcionadas a esses níveis por meio de instrumentos de avaliação para a educação básica (Saeb, Enem, Ideb, Prova Brasil) e pela criação do sistema nacional de avaliação da educação superior (Sinaes), além daqueles específicos para o sistema de avaliação da pós-graduação e da pesquisa (BRASIL, 2010, p. 54).

O texto do documento propõe, então, uma visão ampla de avaliação, de modo a coadunar-se com os fins da educação, com o conceito de ser humano e de diversidade e com o projeto de sociedade, mas de contrapor-se aos modelos atuais, cuja centralidade resulta em controle e competição institucional. O texto aponta:

> [...] a necessidade de novos marcos para os processos avaliativos, incluindo sua conexão à educação básica e superior, aos sistemas de ensino e, sobretudo, assentando-os em uma visão formativa, que considere os diferentes espaços e atores, envolvendo o desenvolvimento institucional e profissional (BRASIL, 2010, p. 53).

Para isso, foi recomendada a criação de um sistema nacional de avaliação, o que implicaria na articulação dos entes federados, a partir da implementação de uma política voltada para a melhoria da educação. Em termos práticos, a defesa é de uma avaliação que considere:

> [...] o rendimento escolar, mas, também, situar as outras variáveis que contribuem para a aprendizagem, tais como: os impactos da desigualdade social e regional na efetivação e consolidação das práticas pedagógicas, os contextos culturais nos quais se realizam os processos de ensino e aprendizagem; a qualificação, os salários e a carreira dos/das professores/as; as condições físicas e de equipamentos das instituições; o tempo

> de permanência do/da estudante na instituição; a gestão democrática; os projetos político-pedagógicos e planos de desenvolvimento institucionais construídos coletivamente; o atendimento extraturno aos/às estudantes que necessitam de maior apoio; e o número de estudantes por professor/a em sala de aula, dentre outros, na educação básica e superior, pública e privada (BRASIL, 2010, p. 53).

Educação de qualidade e avaliação no PNE (2014-2024)

O documento final da Conae teve como finalidade ser a base da proposta do próximo Plano Nacional de Educação, sendo esse o principal motivo da sua realização. Esse documento retratou bem a preocupação dos participantes da Conae com relação à qualidade da educação e avançou no entendimento da avaliação, vista em uma concepção mais ampla e não restrita ao rendimento escolar. No entanto, o texto do Projeto de Lei nº 8.035, encaminhado pelos técnicos do Ministério da Educação (MEC) e publicado em 20/12/2010, não traduziu as definições contidas no documento da Conae, e o texto final da Lei nº 13.005/2014 representou os embates dos dois documentos que serviram como subsídio para a elaboração do PNE 2014-2024.

Nos embates travados na tramitação e aprovação do PNE, ao serem definidas as metas até 2024 para a educação, buscou-se garantir, no escopo legal, a relação entre avaliação e qualidade, atrelada à questão da expansão como garantia do direito à educação para todos. Como aponta Casali (2011), quantidade e qualidade são dimensões que se implicam reciprocamente, e a separação dessas dimensões é uma distorção ontológica. Embora os embates travados no Congresso Nacional tenham mostrado que impera no senso comum e na cultura do mercado a acumulação material e o hiperconsumo, ou seja, a sobrevalorização da quantidade em detrimento da qualidade, havia grupos com clara noção de que:

> [...] quantidade refere-se à extensão e qualidade refere-se ao modo. E mais: não há qualidade sem quantidade, nem vice-versa. Cotidianamente lidamos com ambos os conceitos e, quase sempre, de modo articulado: de tudo o que é bom (qualidade) desejamos mais (quantidade) e melhor (qualidade). No campo da educação, particularmente,

> ambos os conceitos são indissociáveis, mas a quantidade é, ela própria, sempre, parte da substância da qualidade, porque a educação é um direito universal, que deve ser estendido (extensão = quantidade) a todos. O filósofo italiano Antonio Gramsci demarcou com notável clareza essa questão: "Dado que não pode existir quantidade sem qualidade e qualidade sem quantidade (economia sem cultura, atividade prática sem inteligência e vice-versa), toda contraposição dos dois termos é, racionalmente, um contrassenso" (CASALI, 2011, p. 17).

Havia, por parte do grupo da Conae e daqueles que enfrentaram os debates no Congresso, o entendimento de que o plano precisava ser um todo orgânico, no sentido de que o alcance de uma meta, com a qualidade e quantidade desejadas, dependia do alcance das demais. Assim, a luta travada pela garantia do direito a educação de qualidade perpassou a educação básica, a superior, a formação e a valorização dos profissionais da educação, a gestão e o financiamento da educação como um todo.

Dessa forma, a Lei nº 13.005/2014 define no seu art. 5º que a execução do PNE, assim como o cumprimento de suas metas, será objeto de monitoramento contínuo e de avaliações periódicas. O § 2º do citado artigo estabelece que:

> § 2º A cada 2 (dois) anos, ao longo do período de vigência deste PNE, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – INEP publicará estudos para aferir a evolução no cumprimento das metas estabelecidas no Anexo desta Lei, com informações organizadas por ente federado e consolidadas em âmbito nacional, tendo como referência os estudos e as pesquisas de que trata o art. 4º, sem prejuízo de outras fontes e informações relevantes (BRASIL, 2014).

O art. 11 aponta para a criação do Sistema Nacional de Avaliação da Educação Básica (Sinaeb), que deve ser coordenado pela União, em colaboração com os estados, o distrito federal e os municípios, com a finalidade de se constituir como fonte de informação para a avaliação da qualidade da educação básica e para a orientação das políticas públicas desse nível de ensino. O §1º desse artigo aponta que o sistema de avaliação em tela produzirá, no máximo a cada dois anos:

> I) indicadores de rendimento escolar, referentes ao desempenho dos (as) estudantes apurado em exames nacionais de avaliação, com participação de pelo menos 80% (oitenta por cento) dos (as) alunos (as) de cada ano escolar periodicamente avaliado em cada escola, e aos dados pertinentes apurados pelo censo escolar da educação básica;
>
> II) indicadores de avaliação institucional, relativos a características como o perfil do alunado e do corpo dos (as) profissionais da educação, as relações entre dimensão do corpo docente, do corpo técnico e do corpo discente, a infraestrutura das escolas, os recursos pedagógicos disponíveis e os processos da gestão, entre outras relevantes (BRASIL, 2014).

A Lei do PNE, em que pese seus limites, ao apontar para a criação do Sinaeb, aponta também para uma concepção de avaliação que não está voltada apenas aos resultados da aprendizagem dos alunos, mas ao próprio processo de aprendizagem e às condições em que ela acontece, tendo em vista que esses elementos são fundamentais para o alcance da formação desejada. No corpo da lei, apesar de não aparecer todos os aspectos intra e extraescolares, há a intenção de que esses elementos devam compor o instrumento de criação do sistema. Além disso, nas metas que compõem o anexo da lei, é possível encontrar, na perspectiva da qualidade e dos meios de alcance, tanto elementos limitantes quanto avanços.

Ligada à meta 1, que se refere à expansão do atendimento a crianças de zero a três anos e da universalização da educação infantil para crianças de quatro e cinco anos de idade, a estratégia 1.1 aponta que essa expansão deve se dar em regime de colaboração entre os três entes federados (União, estados, municípios e distrito federal), segundo o padrão nacional de qualidade e considerando-se as peculiaridades locais.

Na estratégia 1.6, a articulação entre avaliação e qualidade aparece quando se aponta que deve ser implantada:

> 1.6) [...] até o segundo ano de vigência deste PNE, avaliação da educação infantil, a ser realizada a cada 2 (dois) anos, com base em parâmetros nacionais de qualidade, a fim de aferir a infraestrutura física, o quadro de pessoal, as condições de gestão, os recursos pedagógicos, a situação de acessibilidade, entre outros indicadores relevantes (BRASIL, 2014).

As metas 2 e 3, que tratam do atendimento ao ensino fundamental e médio, respectivamente, também abordam a questão da qualidade. A meta 4, em sua estratégia 4.14, estabelece que, até o segundo ano de vigência do PNE, tem que se definir "indicadores de qualidade e política de avaliação e supervisão para o funcionamento de instituições públicas e privadas que prestam atendimento a alunos com deficiência, transtornos globais do desenvolvimento e altas habilidades ou superdotação" (BRASIL, 2014).

A meta 7 do PNE é destinada à avaliação e vincula a concepção de qualidade da educação às médias do Índice de Desenvolvimento da Educação Básica (Ideb), em uma perspectiva reducionista e parcial, e aponta como princípio o fomento à qualidade da educação básica em todas as etapas e modalidades. Assim, é possível encontrar, em algumas das suas estratégias, uma lógica de qualidade mais abrangente do que a apresentada na meta. As estratégias 7.4 e 7.5 trazem perspectivas de avaliação que vão além do olhar para os resultados, apontando mecanismos de avaliação voltados para a melhoria da educação ofertada, como:

> 7.4) induzir processo contínuo de autoavaliação das escolas de educação básica, por meio da constituição de instrumentos de avaliação que orientem as dimensões a serem fortalecidas, destacando-se a elaboração de planejamento estratégico, a melhoria contínua da qualidade educacional, a formação continuada dos(as) profissionais da educação e o aprimoramento da gestão democrática;
>
> 7.5) formalizar e executar os planos de ações articuladas dando cumprimento às metas de qualidade estabelecidas para a educação básica pública e às estratégias de apoio técnico e financeiro voltadas à melhoria da gestão educacional, à formação de professores e professoras e profissionais de serviços e apoio escolares, à ampliação e ao desenvolvimento de recursos pedagógicos e à melhoria e expansão da infraestrutura física da rede escolar (BRASIL, 2014).

Outras estratégias dessa meta apontam o desenvolvimento de indicadores específicos de avaliação da qualidade da educação especial, bem como da qualidade da educação bilíngue para surdos (7.8) e, ainda, o estabelecimento de forma articulada de parâmetros mínimos de qualidade da educação básica "a serem utilizados como referência para infraestrutura

das escolas, recursos pedagógicos, entre outros insumos relevantes, bem como instrumento para adoção de medidas para a melhoria da qualidade do ensino" (BRASIL, 2014).

Nas estratégias 7.29 e 7.31, há indicativos de articulação dos programas da área da educação com os de outras áreas, como saúde, trabalho e emprego, assistência social, esporte e cultura, para possibilitar a criação de uma rede de apoio integral às famílias, além de realizar ações efetivas de prevenção, atenção e atendimento à saúde e à integridade física, mental e emocional dos profissionais da educação, como condição para a melhoria da qualidade educacional.

No campo da educação profissional, além da meta 11, que trata da expansão da oferta de educação profissional com qualidade, a estratégia 11.8 aponta para a institucionalização de um "sistema de avaliação da qualidade da educação profissional técnica de nível médio das redes escolares públicas e privadas" (BRASIL, 2014).

A meta 12, que versa sobre a elevação das taxas bruta e líquida de matrícula na educação superior, aponta que é preciso assegurar a qualidade da oferta. Na estratégia 12.14, é apontado o mapeamento da demanda, além do fomento à formação de pessoal de nível superior para o desenvolvimento do País, a inovação tecnológica e a melhoria da qualidade da educação básica. Na estratégia 12.19, o plano indica a reestruturação, com ênfase na qualidade dos procedimentos de avaliação, regulação e supervisão, dos "processos de autorização de cursos e instituições, de reconhecimento ou renovação de reconhecimento de cursos superiores e de credenciamento ou recredenciamento de instituições, no âmbito do sistema federal de ensino" (BRASIL, 2014).

Assim como a educação básica, a educação superior também teve, não sem muita resistência dos empresários da educação, sua meta voltada para a qualidade da oferta. A meta 13 versa sobre a elevação da qualidade da educação superior e, para isso, define que deve haver a ampliação da "proporção de mestres e doutores do corpo docente em efetivo exercício no conjunto do sistema de educação superior para 75% (setenta e cinco por cento), sendo, do total, no mínimo, 35% (trinta e cinco por cento) doutores" (BRASIL, 2014). As estratégias de implementação da meta apontam para

a melhoria da qualidade dos cursos de pedagogia e licenciaturas, para a aplicação de instrumento próprio de avaliação integrados às demandas e necessidades das redes de educação básica, além de indicar a elevação do "padrão de qualidade das universidades, direcionando sua atividade, de modo que realizem, efetivamente, pesquisa institucionalizada, articulada a programas de pós-graduação *stricto sensu*" (BRASIL, 2014).

Por fim, a meta 20 versa sobre as condições de viabilização das outras 19. Essa meta estabeleceu o percentual do Produto Interno Bruto do país como mínimo necessário para a garantia do direito à educação, tendo como parâmetro a educação de qualidade. Ao definir que o investimento público em educação pública deveria alcançar, no mínimo, 7% do PIB do país no quinto ano de vigência da lei e o equivalente a 10% ao final do decênio, o plano visou à inversão da lógica de financiamento presente no País até os dias atuais.

Atualmente, o parâmetro de financiamento é o percentual constitucional de, no mínimo, 18% para a União e de 25% para estados, distrito federal e municípios da receita resultante de impostos. O PNE, ao tomar como referência o Custo Aluno-Qualidade inicial – CAQi e o Custo Aluno-Qualidade – CAQ como parâmetro de financiamento, avançou significativamente, tendo em vista que esses dispositivos têm como referência o conjunto de padrões mínimos estabelecidos na legislação educacional para garantir a qualidade do processo de ensino-aprendizagem. Assim, a estratégia 20.7, ao definir a implementação do CAQ, toma:

> [...] como parâmetro para o financiamento da educação de todas etapas e modalidades da educação básica, a partir do cálculo e do acompanhamento regular dos indicadores de gastos educacionais com investimentos em qualificação e remuneração do pessoal docente e dos demais profissionais da educação pública, em aquisição, manutenção, construção e conservação de instalações e equipamentos necessários ao ensino e em aquisição de material didático-escolar, alimentação e transporte escolar (BRASIL, 2014).

Dentre as diversas estratégias de implementação do PNE que fazem referência à qualidade, destaca-se, a seguir, a tentativa de implementação do Sinaeb.

Sinaeb

Sobre a criação do Sinaeb, o documento da Conae aponta que, em termos objetivos:

> o sistema de avaliação deve ser capaz de identificar os desafios institucionais de infraestrutura dos sistemas de educação (tais como situação do prédio, existência de biblioteca e equipamentos, recursos pedagógicos e midiáticos, condições de trabalho dos/das profissionais de educação, dentre outros) e aferir o processo de democratização nas escolas, utilizando os indicadores de avaliação existentes para garantir a melhoria do trabalho escolar, bem como o aperfeiçoamento do senso crítico do/da estudante (BRASIL, 2010, p. 55).

Seguindo essa orientação, o Inep, por meio da Diretoria de Avaliação da Educação Básica (Daeb), articulou discussões que embasaram a portaria de criação do Sinaeb. Para essa discussão, partiu-se da compreensão de que havia a necessidade de definição de processos avaliativos mais amplos, participativos e diversificados, os quais pudessem oferecer maiores subsídios para a formulação e a melhoria de políticas que fomentassem o desenvolvimento de projetos educativos mais inclusivos e equitativos e que contribuíssem para o aprimoramento das demandas sociais pelo direito à educação. Foram sinalizados alguns aspectos que influenciam o sucesso escolar e que deveriam ser contemplados nesse sistema, a saber: os impactos das desigualdades sociais e regionais nas práticas pedagógicas; os contextos culturais nos quais se realizam os processos de ensino e aprendizagem; a qualificação, os salários e a carreira dos profissionais da educação; as condições físicas e os equipamentos das instituições educativas; o tempo diário de permanência do estudante na instituição; a gestão democrática; os projetos político-pedagógicos e os planos de desenvolvimento institucionais construídos coletivamente; o atendimento extra turno aos/às estudantes e o número de estudantes por professor na escola em todos os níveis, etapas e modalidades, nas esferas pública ou privada, entre outros.

Assim, a Portaria MEC nº 369, de 5 de maio de 2016, no seu art. 1º, instituiu o Sinaeb:

> [...] com o objetivo de assegurar o processo nacional de avaliação da educação básica em todas as etapas e modalidades, considerando suas múltiplas dimensões, na perspectiva de garantir a universalização do atendimento escolar, por meio de uma educação de qualidade e democrática, a valorização dos profissionais da educação e a superação das desigualdades educacionais (BRASIL, 2016).

O § 1º do referido artigo define que o Sinaeb será vinculado ao Sistema Nacional de Educação e coordenado pela União em colaboração com os demais entes federados, constituindo-se fonte de informação para a avaliação da qualidade da educação básica e para a orientação das políticas públicas desse nível de ensino, com base em:

> I) indicadores de rendimento escolar, referentes ao desempenho dos estudantes apurado em exames nacionais de avaliação e aos dados pertinentes apurados pelo censo escolar da educação básica; e
>
> II) indicadores de avaliação institucional concernentes a características como o perfil do alunado e do corpo dos profissionais da educação, as relações entre dimensão do corpo docente, do corpo técnico e do corpo discente, a infraestrutura física, as condições de gestão, os recursos pedagógicos, a situação de acessibilidade, autoavaliação, entre outros indicadores contextuais relevantes, além de fornecer subsídios aos sistemas de ensino para a construção de políticas públicas que possibilitem melhoria na qualidade da educação básica – em todas as suas etapas e modalidade (BRASIL, 2016).

A fim de manter articulação com as metas do PNE, o Sinaeb também se propôs a subsidiar o monitoramento da estratégia 20.10, a qual define que caberá à União complementar os recursos financeiros a todos os estados, ao distrito federal e aos municípios que não conseguirem atingir o valor do CAQi e, posteriormente, do CAQ. Desse modo, o § 2º do art. 1º define que caberá ao Sinaeb produzir:

> [...] indicadores de qualidade das condições de oferta para orientar a ação redistributiva e supletiva, técnica e financeira, do orçamento da União com relação aos Estados, Distrito Federal e Municípios e dos orçamentos dos Estados com relação aos seus Municípios, sendo uma referência para

> a definição do Custo Aluno Qualidade Inicial – CAQi e do Custo Aluno Qualidade – CAQ (BRASIL, 2016).

O órgão também avançou na criação do Comitê de Governança do Sinaeb, com o objetivo de propor, acompanhar e supervisionar a implantação e o desenvolvimento do sistema. Esse comitê seria composto por representantes das seguintes entidades (BRASIL, 2016):

III) Inep;

IV) Secretaria de Educação Básica (SEB-MEC);

V) Secretaria de Articulação com os Sistemas de Ensino (Sase-MEC);

VI) Conselho Nacional de Educação (CNE);

VII) Associação Nacional de Pós-Graduação e Pesquisa em Educação (Anped);

VIII) Confederação Nacional dos Trabalhadores da Educação (CNTE);

IX) União Nacional dos Dirigentes Municipais de Educação (Undime);

X) Conselho Nacional de Secretários de Educação;

XI) Associação Nacional de Política e Administração da Educação (Anpae);

XII) Fórum Nacional de Educação (FNE); e

XIII) Campanha Nacional pelo Direito à Educação.

Outro avanço que pode ser destacado foi a definição, no art. 10, de que o Sinaeb produziria indicadores de qualidade para as diretrizes e as dimensões da avaliação da educação básica, que terão metodologia de coleta, cálculo e divulgação estabelecida em portaria específica do Inep, conforme o comitê de governança do sistema. As diretrizes e as dimensões podem ser vistas no quadro que constitui o anexo da portaria (quadro 1).

Quadro 1 Dimensões e diretrizes de avaliação da educação básica

DIRETRIZ	DIMENSÃO
Universalização do atendimento escolar	Acesso e permanência
	Trajetória
	Infraestrutura
Melhoria da qualidade do aprendizado	Aprendizagens
	Práticas pedagógicas
	Ambiente educativo
	Formação para o trabalho e cidadania
Valorização dos profissionais da educação	Formação inicial e continuada
	Carreira e remuneração
	Satisfação profissional
Gestão democrática	Financiamento
	Planejamento e gestão
	Participação
Superação das desigualdades educacionais	Inclusão e equidade
	Direitos humanos, diversidade e diferença
	Contexto socioeconômico e espacial
	Intersetorialidade e sustentabilidade

Fonte: Portaria MEC nº 369 de 5 de maio de 2016.

Em que pese a ideia de pensar uma perspectiva mais ampla de avaliação, articulada às diretrizes do PNE, a portaria não desvincula o Sinaeb das avaliações existentes: *i*) Avaliação Nacional da Educação Infantil (Anei) – em implementação; *ii*) Provinha Brasil; *iii*) Avaliação Nacional de Alfabetização (Ana); *iv*) Avaliação Nacional de Educação Básica (Aneb); *v*) Avaliação Nacional do Rendimento Escolar (Anresc) – Prova Brasil; e *vi*) Programa Internacional de Avaliação de Alunos (Pisa). Foram propostas: a criação de novos instrumentos para avaliar as distintas modalidades da educação básica; e a progressiva ampliação da participação da rede privada na ANA e na Anresc – Prova Brasil.

Além disso, o Sinaeb propõe a produção de três indicadores: *i*) indicadores de rendimento escolar; *ii*) indicadores de avaliação institucional; e *iii*) Índice de Diferença do Desempenho esperado e verificado (IDD) dos estudantes

da educação básica, que será agregado ao já existente Ideb. Sobre o IDD, alerta-se para o uso de seu resultado, que pode representar um mecanismo de responsabilização, sobretudo dos professores, ao ser analisado o "valor agregado" do impacto da ação pedagógica da escola ou do professor no sucesso escolar do aluno.

Um fato determinante na descontinuidade da implementação do Sinaeb foi o processo de impedimento da presidente da República, Dilma Rousseff, e a transferência do seu cargo para seu vice, Michel Temer, situação que desencadeou mudanças nas pastas dos ministérios, como também nas presidências dos órgãos e autarquias a eles ligados. No MEC, Mendonça Filho, ao se tornar o titular da pasta, revogou medidas tomadas pelo seu antecessor. No dia 26 de agosto de 2016, foi emitida a Portaria MEC nº 981, a qual revogou a Portaria MEC nº 369, de 5 de maio de 2016, que instituiu o Sinaeb.

A revogação do Sinaeb foi amplamente criticada pelos grupos que defendem a garantia do direito à educação de qualidade, tendo em vista que este era um sistema que previa uma avaliação mais próxima do que se almeja em uma perspectiva mais avançada de educação. A Campanha Nacional pelo Direito à Educação[3], uma das propositoras da criação desse sistema e defensora da sua aprovação na lei do PNE, assim se posicionou:

> A rede da Campanha Nacional pelo Direito à Educação repudia a revogação da Portaria nº 369 de 5 de maio de 2016, dedicada a regulamentar o Sistema Nacional de Avaliação da Educação Básica (Sinaeb). [...] o Sinaeb é um instrumento legal destinado a qualificar a avaliação da educação básica, tornando-a capaz de auxiliar verdadeiramente o aprimoramento das políticas educacionais em suas diferentes dimensões, inclusive fazendo melhor uso dos mecanismos avaliativos já existentes e fomentando uma nova cultura avaliativa na educação, além de criar outros extremamente necessários (CAMPANHA NACIONAL PELO DIREITO À EDUCAÇÃO, 2016).

3 A Campanha Nacional pelo Direito à Educação é uma articulação ampla e plural no campo da educação no Brasil, constitui-se como uma rede que articula centenas de grupos e entidades distribuídas por todo o País, incluindo comunidades escolares, movimentos sociais, sindicatos, organizações não-governamentais nacionais e internacionais, fundações, grupos universitários, estudantis, juvenis e comunitários, além de milhares de cidadãos que acreditam na construção de um país justo e sustentável por meio da oferta de uma educação pública de qualidade.

A Portaria nº 981/2016 justifica a revogação considerando as revisões da Base Nacional Curricular Comum, ainda em curso, cujas orientações e recomendações devem nortear a instituição do Sinaeb. No entanto, essa justificativa é frágil, posto que o Sinaeb deve articular a avaliação, conforme orienta o PNE, e não se vincular à questão curricular.

Outro ponto de destaque é o estabelecido no art. 2º, que determina a manutenção das avaliações da educação básica já realizadas pelo Inep. Nessa direção, pretende-se, em vez de um sistema nacional de avaliação da educação básica, ter um conjunto desarticulado de instrumentos de avaliação que, por um lado, traduz concepções e práticas avaliativas de responsabilização, centradas quase exclusivamente no desempenho dos estudantes em testes em larga escala, e por outro lado, não dão conta de avaliar a educação na perspectiva apontada pelo documento da Conae e pelo PNE.

Considerações finais

Este artigo analisou o tema da avaliação no PNE, buscando compreender seu conceito no plano e sua estreita relação com a concepção de educação apresentada. Concluiu-se que a concepção de educação no PNE se mantém vinculada ao dispositivo constitucional da educação como direito de todos e com garantia da qualidade de sua oferta.

Com relação à garantia constitucional de qualidade na educação, concluiu-se, também, que a avaliação no PNE se assume de forma mais sistêmica e mais ampla, servindo como subsídio ao alcance das metas propostas, à formulação e ao desenvolvimento de políticas educacionais apesar da vinculação aos resultados do Ideb.

No que se refere ao Sinaeb, verificou-se que o desenho proposto representou avanços quanto à concepção de avaliação existente, quanto à ampliação do entendimento e das dimensões que passam a compor essa avaliação e, ainda, quanto à ideia de um sistema nacional de avaliação articulado com o PNE.

Disso, compreende-se que as perspectivas do PNE propuseram a avaliação

em uma lógica que visa a garantia do direito a educação de qualidade. No entanto, as políticas efetivadas até o momento não foram suficientes para a implementação dessa lógica, sobretudo pelo movimento de descontinuidade política vivenciado no Brasil. Ademais, percebemos que as concepções de avaliação presentes nas políticas de avaliação no Brasil têm seguido dois modelos: um com base nos resultados de alunos em avaliações em larga escala e na produção de *rankings* entre instituições; e outro com base na compreensão do processo educativo como um todo e na construção de elementos que levem à melhoria dos processos e ao sucesso escolar.

Segundo Chizzotti (2016), a avaliação é um meio histórico de qualificar a educação; visa, essencialmente, garantir o direito inalienável de aprender e não se reduz à quantificação de resultados mensurados de respostas esperadas com base no conteúdo ministrado. Nesse sentido, "as políticas públicas de avaliação precisam incentivar a cultura da avaliação formativa como meio de garantir a avaliação justa dos esforços de todos os que têm o direito de aprender na educação escolar" (CHIZZOTTI, 2016, p. 572).

Diante dos limites, dos avanços e dos recuos apontados até aqui, assim como das indefinições a respeito do que os tempos futuros nos revelarão sobre a garantia de uma educação de qualidade, cabe-nos pautar a seguinte reflexão.

> Qualidade da educação também: em parte é intangível. Não se mede, não se controla; entretanto, é factível. Não se deve confundir o factível com o controlável. Há muitas experiências na vida que são factíveis, mesmo não sendo mensuráveis, tangíveis, controláveis. A intangibilidade da qualidade traz o risco de esvaziar o discurso acerca da qualidade, alegando-se a "impossibilidade de resolver definitivamente a questão". É fato: é impossível resolver definitivamente essa questão, porque ela é interminável. Mas assim também são a ciência, a arte, a sabedoria, o amor, o desenvolvimento dos talentos etc. e nem por isso deixamos de almejá-los (CASALI, 2011, p. 16).

Referências

BRASIL. Ministério da Educação. Documento Final da Conferência Nacional de Educação, de 2010. Brasília, 2010. Disponível em: <http://portal.mec.gov.br/arquivos/pdf/conae/documento_referencia.pdf>. Acesso em: mar. 2017.

BRASIL. Lei nº 13.005, de 25 de junho de 2014. Aprova o Plano Nacional de Educação e dá outras providências. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l13005.htm>. Acesso em: out. 2016.

BRASIL. Ministério da Educação. Portaria nº 369, de 5 de maio de 2016. Institui o Sistema Nacional de Avaliação da Educação Básica. Disponível em: <http://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?jornal=1&pagina=26&data=06/05/2016>. Acesso em: mar. 2017.

BRASIL. Portaria n° 981, de 26 de agosto de 2016. Revoga a Portaria MEC nº 369, de 5 de maio de 2016 e dá outras providências. Disponível em: <http://www.lex.com.br/legis_27180245_portaria_n_981_de_25_de_agosto_de_2016.aspx>. Acesso em: mar. 2017.

DOURADO, L. F.; OLIVEIRA, J. F.; SANTOS, C. A. A qualidade da educação: conceitos e definições. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2007. 65 p. (Série documental. Textos para discussão). Disponível em: <http://www.publicacoes.inep.gov.br/portal/download/521>. Acesso em: mar. 2017.

CAMPANHA NACIONAL PELO DIREITO À EDICAÇÃO. *Posicionamento Público*: MEC revoga novo sistema para avaliação da educação básica previsto no PNE. São Paulo, 1º set. 2016. Disponível em: <http://campanha.org.br/avaliacao/posicionamento-publico-mec-revoga-novo-sistema-para-avaliacao-da-educacao-basica-previsto-no-pne/>. Acesso em: mar. 2017.

CASALI, A. Fundamentos para uma avaliação educativa. In: CAPPELLETTI, I. F. Avaliação da aprendizagem: discussão de caminhos. São Paulo: Editora Articulação Universidade/Escola, 2007.

CASALI, A. O que é educação de qualidade? In: MANHAS, C. (org.). Quanto Custa Universalizar o Direito à Educação? Brasília: Instituto de Estudos Socioeconômicos, 2011. Disponível em: <http://www.inesc.org.br/biblioteca/textos/livros/quanto-custa-universalizar-o-direito-a-educacao>. Acesso em: mar. 2017.

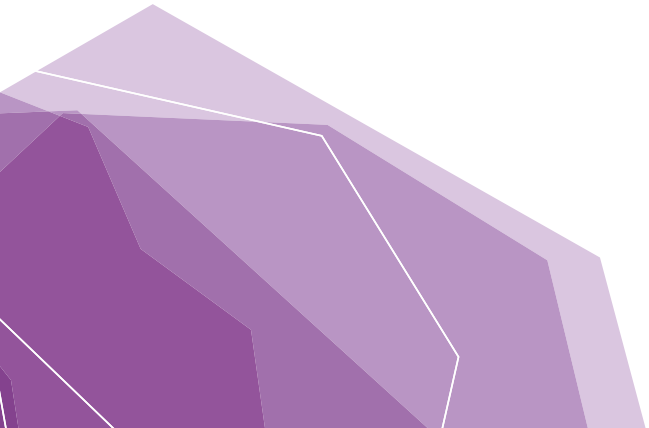
CHIZZOTTI, A. Políticas públicas: direito de aprender e avaliação formativa. Práxis Educativa, Ponta Grossa, v. 11, n. 3. p. 561-576, set./dez. 2016. Disponível em: <http://www.revistas2.uepg.br/index.php/praxiseducativa>. Acesso em: mar. 2017.

Catarina de Almeida Santos
Doutora em Educação pela Universidade de São Paulo
Professora da Universidade de Brasília
cdealmeidasantos@gmail.com

Danielle Xabregas Pamplona Nogueira
Doutora em Educação pela Universidade de Brasília
Professora da Universidade de Brasília
danielle.pamplona@gmail.com

RESENHA

review



CONSTRUINDO TESTES: COMO ELABORAR E VALIDAR ITENS DE MÚLTIPLA ESCOLHA

DEVELOPING TESTS: HOW TO CREATE AND VALIDATE MULTIPLE CHOICE ITEMS

CONSTRUYENDO TESTS: CÓMO ELABORAR Y VALIDAR ÍTENS DE OPCIÓN MÚLTIPLE

Alessandra Ramos de Oliveira Harden

Haladyna, T. M. *Developing and validating multiple-choice test items.*

Publicado pela primeira vez em 1994 e ainda, para nosso azar, sem tradução para o português, *Developing and validating multiple-choice test items,* de Thomas Haladyna pode ser considerado como leitura fundamental para qualquer um que trabalhe com avaliação educacional. De fato, o autor já esclarece, logo na introdução, que escreveu para "aqueles seriamente interessados na elaboração de itens para a avaliação de desempenho" (p. viii).[1] Esses "interessados" são colocados por Haladyna em dois grupos: o primeiro é formado por estudantes universitários das áreas de mensuração educacional, que têm na leitura do livro a oportunidade de compreender melhor as duas fases da elaboração de testes, que seriam a elaboração ou construção propriamente dita do teste e a validação das respostas dadas. O segundo grupo é constituído por profissionais que trabalham diretamente com a elaboração de itens, para quem certamente Haladyna apresenta amplo material para ajudá-los a aprimorar sua prática. Ou seja, esse é um livro também para todos os professores.

Sem dúvida, o que dá o tom ao livro é a familiaridade do autor com o tema e a propriedade com a qual lida com as muitas questões ligadas à avaliação de desempenho. Isso não é surpresa para quem sabe da vasta experiência do

1 Todas as traduções de trechos do livro aqui citados são de minha autoria.

autor. Professor emérito da Arizona State University, Haladyna tem uma longa lista de publicações que demonstram sua dedicação à psicologia educacional, à psicometria e a outros tópicos ligados à elaboração de testes (para aplicação tanto em larga escala quanto para turmas pequenas). O saber adquirido com os anos de trabalho é mencionado como uma das justificativas para a terceira edição de *Developing and validating multiple-choice test items*:

> [...] meus mais de 30 anos de experiência com planejamento, aplicação e avaliação de sistemas de testagem, e com ensino no nível fundamental, na graduação e na pós-graduação me ajudaram a entender melhor o processo de elaboração de itens, as vantagens trazidas por itens bem construídos e a importância da validação das respostas dadas (p. viii).

Essa vivência de Haladyna em ambientes ligados ao ensino se manifesta não apenas na riqueza de informação e no cuidado com o detalhe, mas também na qualidade didática do próprio texto. As quase 300 páginas são escritas de maneira clara e organizada e formam capítulos bem concatenados, ao longo dos quais Haladyna vai gradual e pacientemente guiando o leitor pelo universo das avaliações com base em itens de múltipla escolha (ME). Pouco a pouco, o conteúdo denso é apresentado, e até tópicos mais espinhosos começam a fazer mais sentido. A receita do sucesso do professor: ele define até mesmo os conceitos mais básicos e adiciona ao texto tabelas e quadros sinópticos, além de presentear o leitor com pequenas introduções e sumários em cada capítulo.

É assim, portanto, que está estruturado cada um dos onze capítulos do livro, que são precedidos por uma introdução. Nela, Haladyna explica que a terceira edição é uma versão revisada e melhorada das outras duas edições e expõe os motivos para tanto: o contínuo interesse dos leitores, a necessidade de ampliar a pesquisa sobre a construção de itens, o desenvolvimento na construção de itens de ME e a já mencionada convicção do autor nos benefícios de itens bem feitos e na relevância da validação. Encontram-se também nesse prefácio breves considerações sobre o que Haladyna considera limitações do livro, sobre o então atual estado dos testes de ME, além da indicação do tipo de leitor a quem o livro é dirigido, conforme citado acima, e uma apresentação do modo como o livro está organizado, com os capítulos divididos em quatro partes.

Um ponto que chama atenção ainda nesse trecho preambular é que o professor Haladyna critica a ênfase dada ao uso da ME, sugerindo que possa ter sido demasiada. No entanto, ele reconhece o papel crucial desse tipo de item tanto nos Estados Unidos, país cuja realidade guia suas considerações, quanto em outras partes do mundo. De fato, é inegável o amplo e generalizado uso da ME, que vem sendo empregada com objetivos diversos, como seleção de estudantes e de profissionais, certificação, concessão de créditos de cursos, atribuição de notas e, claro, avaliação da aprendizagem. Acertadamente, o autor não se limita a nenhum desses usos em particular, mas, ao contrário, declara que uma das premissas de seu livro é o reconhecimento do valor de itens de ME em sala de aula, em avaliações educacionais em larga escala ou em testes de proficiência profissional.

A primeira seção do livro, intitulada "A foundation for multiple-choice testing" (Fundamentos dos testes de ME), é dedicada a apresentar e contextualizar os testes que são o objeto de interesse de Haladyna na obra. São três os capítulos aí agrupados. No primeiro, "The importance of item development for validity", o autor apresenta conceitos básicos para que os leitores possam compreender a extrema relevância da validade na construção de itens. Entende-se validade como "um processo lógico no qual definimos o que estamos medindo, criamos as medidas para fazê-lo, coletamos dados e avaliamos esses dados no que se refere à validade de interpretação de um escore de teste e seu uso subsequente" (p. 3). O mais fundamental dos conceitos é o de item, ou *test item*, e o autor se preocupa a logo defini-lo e caracterizá-lo:

> [...] um item é a unidade básica observável de qualquer teste que geralmente contém uma afirmação que provoca ou exige uma resposta por parte do avaliando. Essa resposta recebe um valor numérico específico, quase sempre 1 quando está correta e 0 quando está errada, mas pode também ser alocada em uma escala de valores indo do baixo ao alto (p. 3).

O capítulo segue com esclarecimentos acerca de noções essenciais sobre as fases de construção de itens de ME. O leitor encontra aí uma lista das tarefas que cabem aos responsáveis pelo instrumento de avaliação, uma espécie de *checklist* das etapas a serem seguidas: "1 – Faça um planejamento para a elaboração dos itens"; 2 – Crie um cronograma para a elaboração dos itens;" e assim por diante (p. 14). Há também uma tabela com as fases do processo de validação dos itens, o qual Haladyna divide em formulação, explicação e validação (p. 18).

O segundo capítulo — "Content and Cognitive Processes" — está ligado à identificação do que se deseja medir com um instrumento de avaliação, ou seja, à definição dos conteúdos disciplinares e dos processos cognitivos que devem guiar a elaboração dos itens. Há uma taxonomia segundo a qual a aprendizagem pode ser compreendida (conhecimento, competências e habilidades), e o texto é enriquecido com copiosos exemplos.

No capítulo três, de título "Item formats", o foco é a diferença entre tipos (ou formatos) de itens de teste, a qual reside no fato de o construto medido ser abstrato ou concreto. O aprendizado definido em termos concretos é designado também como de baixa inferência e aquele definido em termos abstratos é de alta inferência. Haladyna propõe distinções importantes entre os dois tipos de formatos em uma tabela cuja clareza e praticidade a tornam muito útil. Mais adiante no capítulo, a atenção se volta à avaliação do formato de item com base na validade, e o autor expõe seis argumentos de validade que servem para verificar a propriedade do uso dos formatos de ME em contextos específicos.

"Developing MC test items" (Desenvolvimento de itens de ME) é o nome da segunda seção do livro. Nela estão contidos quatro capítulos que formam um verdadeiro guia prático para a construção de itens. O capítulo quatro ("Item formats") consiste em um estudo de oito formatos de ME, com indicações de conteúdos e processos cognitivos para cuja mensuração cada um deles é mais recomendável e discussão de vantagens e desvantagens de seu uso. Novamente, o autor adiciona muitos exemplos.

O capítulo cinco tem o título "Guidelines for developing MC items" e é realmente isso: uma coletânea de 31 diretrizes dirigidas a elaboradores de itens de ME. Abrange desde cuidados gerais a serem observados com relação a conteúdo até aspectos da escrita de enunciados e distratores (por exemplo: cada item deve refletir somente um tópico do conteúdo e somente um processo cognitivo; os distratores devem ser plausíveis etc.). Além disso, o texto traz orientações específicas para cada um dos formatos de itens apresentados no capítulo anterior. Trata-se de material precioso especialmente para os que atuam na revisão técnica e linguística de itens. Sem dúvida, pode ajudar a diminuir o grande número de itens descartados em qualquer processo de elaboração e validação de instrumentos de avaliação.

No capítulo seguinte, o autor demonstra a aplicação do conteúdo tratado até aí, pois expõe o leitor a uma série de exemplos de itens, discutindo-os com o objetivo de evidenciar a variedade de formatos que os itens de ME podem assumir, dependendo da área de conhecimento, do que se deseja medir e, claro, da criatividade dos elaboradores envolvidos.

Já o capítulo sete ("Item generation") é motivado pela necessidade contínua que se tem de gerar itens cada vez melhores. Diante desse desafio, Haladyna argumenta, "qualquer estratégia que aumente a qualidade dos itens e a velocidade de produção é bem-vinda" (p. 148). Nessa linha, o professor comenta brevemente o futuro das teorias de geração de itens para em seguida descrever, exemplificar e avaliar técnicas utilizadas para ajudar elaboradores e diminuir o risco de descarte.

A terceira seção está voltada para a validação de itens. Intitulada "Validity evidence arising from item development and item response validation" (Evidência de validade com base na elaboração do item e na validação da resposta ao item), é composta por três capítulos que versam sobre aspectos do processo de validação da resposta dada ao item.

No capítulo oito ("Validity evidence coming from item development procedures"), o autor recorre mais uma vez a uma tabela, em que lista seis regras ou princípios referentes à qualidade dos itens. Após isso, divide o capítulo em duas subseções: uma dedicada a questões gerais, em que aborda tópicos como a definição de conteúdo, a elaboração de guias para escrita de itens, a seleção e o treinamento de elaboradores, e a segurança; e outra, em que se concentra na revisão dos itens. Nesta, são elencadas oito tipos de revisão que devem ser executadas antes da pré-testagem do material.

No capítulo nove, "Validity evidence coming from statistical study of item responses", Haladyna dá início à discussão do tratamento estatístico das respostas aos itens, revisando aspectos da Teoria Clássica dos Testes e da Teoria de Resposta ao Item. Para tanto, utiliza três abordagens "diferentes, mas complementares" (p. 202) da resposta ao item: a tabular, a gráfica e a estatística. Por fim, propõe orientações para a avaliação de itens com base na combinação do nível de dificuldade com o de discriminação atribuídos a cada item.

No capítulo dez, a preocupação é avaliar quatro pontos da análise de respostas ao item que podem colocar em risco a validade. Esses problemas são abordados individualmente em subcapítulos, em que o autor indica e discute material bibliográfico.

Na última seção da obra, "The future of item development and item response validation" (O futuro da elaboração de itens e da validação da resposta ao item), formada pelo capítulo onze ("New directions in item writing and item response validation"), o autor tece comentários sobre os temas que, na sua opinião, mais podem influenciar o futuro das grandes áreas que são objeto do livro: as políticas educacionais, a noção de validade, o desenvolvimento da psicologia cognitiva e a forma como é definido o resultado da escolaridade e da capacitação profissional. Ele discute ainda novas teorias e direcionamentos referentes à elaboração de itens.

É inegável que *Developing and validating multiple-choice test items* tem muitas e importantes qualidades. A longa lista de referências bibliográficas é mais um ponto que justifica sua classificação como essencial. Há obviamente algumas falhas que devem ser observadas e com as quais os leitores devem tomar cuidado, mas que estão mais ligadas à passagem do tempo e ao desenvolvimento da tecnologia e das pesquisas da área do que a alguma propriedade intrínseca do texto. Se a publicação foi feita em 2004, a leitura exige a complementação das informações com textos mais recentes, especialmente quanto ao emprego de tecnologia digital para a elaboração e aplicação de instrumentos de avaliação e para o tratamento de dados estatísticos. Além disso, o crescimento, no Brasil, das pesquisas voltadas às áreas focadas no livro e o aprimoramento por que passaram nossos sistemas de avaliação também tornam imprescindível o cuidado do leitor para que os ensinamentos do professor Haladyna possam ser relacionados à nossa realidade atual.

Alessandra Ramos de Oliveira Harden
Doutora em Estudos Hispânicos e Lusófonos pela University College Dublin, Irlanda
Professora da Universidade de Brasília
oliveira.ales@gmail.com

ENTREVISTA

interview

PSICOMETRIA NAS AVALIAÇÕES

PSYCHOMETRY IN ASSESSMENT

LA PSICOMETRÍA EN LAS EVALUACIONES

Professor Jacob Arie Laros

Professor Associado do Instituto de Psicologia da Universidade de Brasília. Coordenador do laboratório de métodos e técnicas de avaliação (Meta). Coordenador do Grupo de Trabalho Avaliação Cognitiva e Neuropsicológica da Associação Nacional de Pesquisa e Pós-graduação em Psicologia (ANPEPP). Ph.D. em Psicologia pela University of Groningen.

Examen – Na sua opinião, qual a contribuição da Psicometria para a avaliação educacional?

Jacob – Em alguns artigos sobre o tema existe a lembrança de que, para compreender como a Psicometria pode auxiliar de forma mais efetiva as avaliações educacionais, primeiramente é importante resgatar sua própria definição. Ela não é apenas um conjunto de técnicas e procedimentos para construção e validação de testes, mas bem mais que isso. A Psicometria também é um campo de conhecimento da Psicologia voltado para a obtenção de evidências de construtos psicológicos e modelos teóricos. A psicometria não pode ser confundida com estatística ou com um procedimento técnico para testagem psicológica, mas pode, sim, ser vista como a busca por inovação e por obtenção de evidências que permitam ao psicólogo investigar seus modelos teóricos, compará-los e refutá-los por meio de métodos quantitativos. A Psicometria pode ser uma importante ferramenta para auxiliar no questionamento empírico de seus próprios construtos e objetos de investigação, como é o caso das avaliações educacionais.

Examen – Atualmente a Psicometria é bem utilizada nas avaliações educacionais?

Jacob – Poderia ser mais. Houve um grande avanço nos instrumentos para avaliar proficiência em Matemática e em Português, mas ainda precisamos avançar mais. Por exemplo, no caso dos questionários contextuais, metade das perguntas nem entram de fato na análise fatorial; há problemas na elaboração dos itens – é feita mais de uma pergunta em um mesmo item e se o respondente marca "sim" como resposta, nós não saberemos se a resposta se refere a primeira ou a segunda pergunta. Ainda se cometem erros básicos na construção de medidas educacionais no Brasil. Precisamos entender melhor o construto investigado para depois propor a medida.

Examen – O que falta para aplicarmos mais Psicometria nas avaliações?

Jacob – Arrisco dizer que o que falta talvez seja conhecimento. As pessoas que trabalham com as principais avaliações educacionais no Brasil têm que aprender mais sobre Psicometria para aplicá-la melhor. Falta aprendizagem e talvez também falte tempo. Muitas vezes há pouquíssimo espaço de tempo entre a aplicação de um exame educacional e outro, o que compromete o estudo da metodologia a ser utilizada e a escolha da melhor delimitação do que deve ser investigado nos itens, prejudicando, assim, a construção de medidas educacionais mais eficazes. Outra coisa que falta é valorização de construção de instrumentos psicológicos nas universidades.

Examen – Como o senhor vê a questão da validade na área de educação?

Jacob – A validade é um conceito difícil e que está em constante evolução. Primeiramente, é importante destacar que a fidedignidade e a validade se referem aos escores, não ao teste. Em 1985, a validade era operacionalizada por meio de estudos classificados em três tipos: validade de conteúdo, critério e construto. Depois, em 1999, acrescentou-se a validade consequencial e a validade relacionada com os processos de resposta. Porém, ainda é

comum a utilização do conceito antigo de validade. Precisamos observar a atualização e a discussão do conceito. Falamos sobre evidência de validade e evidências são *ad infinitum* – não têm fim. É necessário sempre continuar a pesquisa sobre a validade do teste e utilizar diferentes formas de investigar essa característica psicométrica (evidências com base: no conteúdo, no processo de resposta, na estrutura interna, nas relações com variáveis externas e nas consequências da testagem). Claro que para testes educacionais o que é mesmo importante é a validade de conteúdo. É preciso observar se os itens realmente atendem o que foi estabelecido na matriz. Além disso, as matrizes de referência também precisam ser revisitadas para que seja observado se estamos mesmo medindo o conteúdo mais importante do domínio.

Examen – As grandes provas aplicadas no Brasil, como o Enem, possuem evidências de validade suficientes?

Jacob – Eu diria que não o suficiente, mas a pergunta é difícil. Podemos ainda melhorar. É importante analisar as diferentes fontes de validade. Não é suficiente realizar somente uma análise. A literatura da área fornece pelos menos cinco fontes de investigação. Assim, não se pode realizar apenas uma análise, de validade de conteúdo, por exemplo, e afirmar que os escores são válidos. É preciso seguir os procedimentos apontados na literatura da área para conseguir embasar as interpretações dos escores para o fim estabelecido.

Examen – Se desenvolvêssemos os estudos sobre as evidências de validade, o que isso acarretaria para a melhoria das avaliações educacionais?

Jacob – O primeiro benefício seria entender melhor as diferentes fontes de validade, o que traria ganhos, principalmente, para os questionários contextuais. Outro aspecto igualmente importante é relativo à formação das pessoas que trabalham com essas questões, que devem ter uma educação continuada no tema. Professores e diretores precisam de capacitação para saber como interpretar os resultados das escolas e, assim, passar a contribuir mais precisamente com melhorias.


Cebraspe
Centro Brasileiro de Pesquisa em Avaliação
e Seleção e de Promoção de Eventos