Exames estandardizados: análise dos modelos e das teorias na produção acadêmica

RODRIGO MARQUES[I]
RONILDO STIEG[II]
WAGNER DOS SANTOS[III]
http://dx.doi.org/10.22347/2175-2753v12i34.2342

Resumo

Analisou-se a constituição do debate sobre os exames estandardizados em diferentes países. Foi utilizado como referencial teórico-metodológico a revisão de literatura em periódicos. As fontes foram mapeadas por meio dos indexadores *SciELO*, *Sibradid*, EBSCO, *Lilac's*, *Latindex*, *Redalyc*, *Elsevier*, *Directory of Open Access Journals*, *Open Edition*, *Dialnet*, visando a extrair os núcleos informativos. Utilizou-se, como descritores, os termos: avaliação em larga escala e exames estandardizados, sem delimitar periodização. Os resultados demonstram que as pesquisas, tomando como referência o modo como os exames estandardizados, têm se constituído em 11 países e evidenciam que a centralidade desses estudos está em discutir o modelo psicométrico e/ou as teorias que fundamentam esses exames.
Palavras-chave: Avaliação em larga escala. Exames estandardizados. Modelos de avaliação.

Submetido em: 23/05/2019
Aprovado em: 22/02/2020

[I] Universidade Federal do Espírito Santo (UFES), Vitória (ES), Brasil; http://orcid.org/0000-0002-8630-2987; e-mail: rodrigo30mar_@hotmail.com.
[II] Universidade Federal do Espírito Santo (UFES), Vitória (ES), Brasil; http://orcid.org/0000-0001-8698-4087; e-mail: roni.stieg@gmail.com.
[III] Universidade Federal do Espírito Santo (UFES), Vitória (ES), Brasil; http://orcid.org/0000-0002-9216-7291; e-mail: wagnercefd@gmaill.com.

Standardized exams: analysis of models and theories in academic production

Abstract

The constitution of the debate on standardized exams in different countries was analyzed. A literature review in journals was used as a theoretical-methodological reference. The sources were mapped using the SciELO, Sibradid, EBSCO, Lilacs, Latindex, Redalyc, Elsevier, Directory of Open Access Journals, Open Edition, Dialnet indexers, in order to extract the information cores. The following terms were used as descriptors: large-scale evaluation and standardized exams, without limiting periodization. The results demonstrate that the research, taking as a reference the way in which standardized examinations, have been constituted in eleven countries, show that the centrality of these studies is in discussing the psychometric model and / or the theories that underlie these examinations.
Keywords: Large-scale evaluation. Standardized exams. Evaluation models.

Exámenes estandarizados: análisis de modelos y teorías en producción académica

Resumen

Se analizó la constitución del debate sobre los exámenes estandarizados en diferentes países. La revisión de la literatura en periódicos se utilizó como marco teórico-metodológico. Las fuentes se mapearon utilizando indexadores SciELO, Sibradid, EBSCO, Lilacs, Latindex, Redalyc, Elsevier, Directory of Open Access Journals, Open Edition, Dialnet, para extraer los núcleos de información. Los siguientes términos se utilizaron como descriptores: evaluación a gran escala y exámenes estandarizados, sin delimitar la periodización. Los resultados muestran que la investigación, tomando como referencia la forma en que se han constituido los exámenes estandarizados en once países, muestran que la centralidad de estos estudios está en discutir el modelo psicométrico y/o las teorías que subyacen a estos exámenes.
Palabras clave: Evaluación a gran escala. Exámenes estandarizados. Modelos de evaluación.

Introdução

Os exames estandardizados, também denominados *testes educacionais padronizados* (THORNDIKE, 1921), *testes de avaliação comportamental* (TYLER, 1949), *avaliação de programas* (WORTHEN; SANDER; FITZPATRICK, 2004) nos Estados Unidos; *avaliação de sistemas educacionais* (GIPPS, 1998) na Inglaterra e País de Gales; *avaliação sistêmica* (LAVASSEUR, 2005) na França; *avaliações do sistema* em Portugal (FERNANDES, 2007; FERREIRA, 2015), no Chile (ARANCIBIA, 1997) e também no Brasil (SILVA, 2010; SUDBRACK; COCCO, 2014), apresentam-se de forma crescente nas políticas educativas contemporâneas nacionais e internacionais[1]. Além disso, esses exames evidenciam um investimento político e financeiro cujo objetivo é direcionar os sistemas educativos.

Entretanto, para delimitação conceitual, é preciso considerar que os *exames estandardizados* se dividem de acordo com sua natureza (nacional ou internacional); seus propósitos (certificação e/ou produção de metadados); e seus efeitos (*moderate stakes* ou *high stakes* – moderados ou de alto risco; *low stakes* – sem efeitos ou com efeitos fracos).

Os *high stakes* são exames de alto risco aplicados em contextos nacionais, cujos resultados são usados para tomar decisões importantes que afetam estudantes, professores, administradores, comunidades, escolas e distritos. Especificamente, fazem parte do desempenho e da política que "[...] vincula a pontuação de um conjunto de testes padronizados à promoção de notas, à conclusão do ensino médio e, em alguns casos, ao professor e ao diretor" (ORFIELD; WALD, 2000, p. 38, tradução nossa).

Já os exames *low stakes*, de acordo com Bauer, Alavarse e Oliveira (2015, p. 1371), são "[...] testes padronizados que não têm consequências sobre a população avaliada, direta (alunos) ou indiretamente (professores, gestores etc.)". Nesse sentido, considerando os propósitos dos exames estandardizados do tipo *low stakes*, identificou-se que eles não oferecem efeitos diretos para o percurso acadêmico dos alunos e para a carreira dos professores, porém fornecem elementos para subsidiar políticas públicas educacionais.

[1] Diante da diversidade de nomenclaturas, assumiu-se, nesta pesquisa, o termo *exames estandardizados*, pois ele é suficiente para atender às diferenças conceituais presentes nos países apresentados.

Entende-se que o uso dos exames estandardizados, em diferentes países, tem gerado problematizações relevantes, ajudando a evidenciar as intencionalidades daquilo que se assume como a principal ferramenta para coleta e construção de bancos de dados sobre determinado sistema educacional. Consequentemente, os dados se revertem em novas diretrizes ou políticas, visando ao desenvolvimento da qualidade do ensino.

Contudo, para o desenvolvimento deste estudo, levantaram-se as seguintes questões: o que são os exames estandardizados e como eles vêm se configurando no debate acadêmico em âmbito mundial? Quais conceituações são utilizadas em diferentes países? Quais são as principais críticas e/ou tensões acenadas por esses estudos?

Para tanto, traçou-se como objetivo analisar como tem se constituído o debate dos exames estandardizados no campo acadêmico em diferentes países a partir da produção científica publicada em periódicos, a fim de produzir um mapa internacional sobre o tema. Nesse caso, não houve um direcionamento na delimitação das fontes pela natureza dos exames estandardizados (nacional ou internacional), mas pelo modo como os autores dos artigos têm delimitado e abordado o tema em diferentes países.

A produção científica indica que o tema tem sido explorado: nos aspectos históricos de determinadas políticas de avaliação (CASTRO, 2000; FERNANDES, 2007; BROOKE, 2008; WERLE; THUM; ANDRADE, 2009; CAICEO ESCUDERO, 2015); nos impactos dos exames estandardizados na avaliação de sala de aula (TORANZOS, 2014; FERREIRA, 2015; BARROS; TAVARES; MASSI, 2009); na avaliação da escrita acadêmica e participação cidadã (NAVARRO; REYES; VERA, 2019); nas suas influências nos currículos e no trabalho docente (FINI, 2009; BONAMINO; SOUZA, 2012; BECKER, 2012; SCHNEIDER, 2013); ou nas críticas às políticas de avaliação (ALTMANN, 2002; ESTEBAN, 2009; ARAÚJO; FERNANDES, 2009; BECKER, 2010; COSGROVE; CARTERIGHT, 2014; SANTOS, 2014), porém não foram mapeados estudos que se dediquem a analisar o que a própria comunidade acadêmica tem discutido sobre o tema.

Métodos e materiais

O estudo é de caráter qualitativo, do tipo revisão bibliográfica em periódicos da educação. Essa metodologia corresponde a um processo de levantamento, descrição e análise de um corpo do conhecimento em busca de resposta a uma pergunta específica. Segundo Morosini e Fernandes (2014), é por meio da identificação, registro e categorização que a revisão bibliográfica permite estabelecer uma análise da produção científica de uma determinada área em um tempo específico.

As fontes foram mapeadas por meio dos indexadores: *Scientific Electronic Library Online* (SciELO*)*, Sistema Brasileiro de Documentação e Informação Desportiva (Sibradid), EBSCO, Literatura Latino-americana e do Caribe em Ciências da Saúde (Lilacs), *Sistema Regional de Información en Línea para Revistas Científicas de América Latina, el Caribe, España y Portugal* (Latindex), *Red de Revistas Cientificas de America Latina y el Caribe, España y Portugal* (Redalyc*)*, Elsevier, *Directory of Open Access Journals*, *Open Edition*, Dialnet. Nesses indexadores, estão reunidas as revistas referentes ao campo que, por sua vez, disponibilizam artigos de onde podemos extrair os núcleos informativos (CATANI, 1996)[2]. A concentração de periódicos científicos de diferentes países, reunidos nesses indexadores, contribui para a divulgação de artigos produzidos por iniciativas particulares, grupos de pesquisa e/ou parcerias internacionais efetuadas por pesquisadores.

Nesse sentido, os indexadores, por meio dos seus núcleos informativos, contribuem para a construção dos discursos e dos *modus operandi* de divulgação das produções intelectuais. Desse modo, foi construída uma "[...] história serial e repertórios analíticos destinados a informar sobre o conteúdo dos periódicos" (CATANI, 1996, p. 118).

A busca dos artigos nos indexadores foi realizada com os descritores em língua inglesa e portuguesa: *large escale evaluation* e *standardized exams* (avaliação em larga escala e exames estandardizados). No Quadro 1, apresentam-se os critérios estabelecidos para definição das fontes.

[2] Segundo Catani (1996), os núcleos formativos são os conteúdos dos artigos e devem ser entendidos como o ponto de partida para o conhecimento em determinado campo de estudos, levando ao centro da discussão sobre determinada temática.

Quadro 1 – Critérios para o levantamento das fontes

Indexadores	SciELO; Sibradid; EBSCO; Lilacs; Latindex; Redalyc; Elsevier; *Directory of Open Access Journals; Open Edition*; Dialnet
Descritores utilizados	*Large scale evaluation; standardized exams;* avaliação em larga escala; e exames estandardizados
Critérios de inclusão para seleção dos artigos	Ter representatividade nas ciências sociais (*social sciences*) Estar presente na lista de indexadores fornecida pelo portal Capes Ter acesso aberto Ter estabelecido, na sua base de dados, critérios de indexação de periódicos reconhecidos internacionalmente pela comunidade científica
Critérios de exclusão dos artigos	Ser de acesso restrito

SciELO: Scientific Electronic Library Online; Sibradid: Sistema Brasileiro de Documentação e Informação Desportiva; Lilacs: Literatura Latino-americana e do Caribe em Ciências da Saúde; Latindex: Sistema Regional de Información en Línea para Revistas Científicas de América Latina, el Caribe, España y Portugal; Redalyc: Red de Revistas Cientificas de America Latina y el Caribe, España y Portugal.
Fonte: Os autores (2019).

O levantamento de dados foi realizado nos meses de fevereiro e março do ano de 2017, sem delimitar periodização. Consistiu, em um primeiro momento, na leitura dos títulos e resumos dos artigos. Nesse processo, foram mapeados 20 estudos, dos quais quatro são nacionais (brasileiros) e 16 internacionais. Depois de constituído o banco de dados (com os arquivos baixados em PDF), realizou-se a leitura do conteúdo dos artigos com o objetivo de elaborar a sua categorização *a posteriori*.

Durante a realização da pesquisa, foi levado em conta a confluência de diversos dados de diferentes origens geográficas que, ao mesmo tempo, foi apontando para novos indícios e conformando um complexo quadro de realidades e possibilidades (GINZBURG, 2002).

Resultados e discussão

A escolha pelo levantamento de dados nos indexadores seguindo os critérios estabelecidos (Quadro 1) permitiu alcançar uma representação em três continentes: Asiático, Americano e Europeu. Outro movimento realizado consistiu na identificação da procedência dos 20 artigos, por periódico e contexto (país), para o estudo do tema exames *estandardizados*, conforme o Quadro 2.

Quadro 2 – Distribuição dos artigos por revista e país em que o tema foi estudado

Revista	País/Contexto de estudo	Artigos
American Evaluation Association	EUA	1
Education Policy Analysis Archives	EUA	1
Educação e Pesquisa	Brasil Finlândia Portugal	2 1 1
Internetional Review of Education	Alemanha	1
Journal of Instructional Development	Canadá	1
Large-Scale Assessments in Education	Alemanha EUA Noruega	1 2 1
Nassp Bulletin	EUA	1
Perspectiva Educacional	México	1
Piscology	China	1
Research in Science Education	Inglaterra	1
Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación	Brasil	1
Revista Lusófona de Educação	Brasil Portugal	1 1
The Asia-Pacific Education Researcher	Hong Kong	1

Fonte: Os autores (2019).

Conforme o Quadro 2, os artigos que tematizam os exames estandardizados têm circulado em 13 periódicos, com destaque para as revistas: *Large-Scale Assessments in Education* (norte-americana) e *Educação e Pesquisa* (brasileira), ambas com quatro publicações, e a *Revista Lusófona de Educação* (portuguesa), com duas publicações. Os outros dez periódicos disponibilizam uma produção que aborda o tema conforme as categorias assumidas neste estudo.

A representação dos artigos em três continentes América (11), Europa (sete) e Ásia (dois) e sua distribuição em 11 países dão visibilidade ao fenômeno da globalização que permeia o tema exames estandardizados, intermediados por órgãos internacionais como: *International Association for the Evaluation of Educational Achivement* (IEA) e *Organización para la Cooperación y el Desarrollo Económicos* (OCDE). Evidencia, ainda, a necessidade de se criar ações e instrumentos efetivos que promovam melhorias na qualidade da educação impulsionadas pela participação desses países em exames estandardizados como: o

Programa Internacional de Avaliação de Estudantes (Pisa)[3] e o *Trends in International Mathematics and Science Study* (TIMMSS)[4].

A partir dos autores que têm estudado o Pisa em diferentes países, é possível identificar interpretações distintas. Na China, Chen e La Torre (2014) indicam que os seus resultados são usados para ilustrar o procedimento sistemático de avaliação. No Brasil, Toffoli, Andrade, Bornia e Quevedo-Camargo (2016) têm se dedicado a descrever as teorias presentes nos dados analisados pelo exame. Em Portugal, Afonso (2009) adverte sobre as consequências do uso do Pisa para a construção de projetos estatísticos com base nos respectivos indicadores como estratégia de viabilização e ampliação de uma agenda globalmente estruturada para educação. Além disso, Correia, Arelaro e Freitas (2015) estabelecem algumas observações críticas relacionadas com a politização dos resultados do exame no país. Na Finlândia, Salokangas e Kauko (2015) discutem a distorção que o Pisa tem trazido, em especial na análise das razões do sucesso de alunos em escolas finlandesas.

De semelhante modo, os resultados do Pisa têm sido utilizados também em outros estudos para analisar os aspectos da avaliação que se realiza na sala de aula (ARAÚJO; TENÓRIO, 2013), na relação com a prática de ensino dos conteúdos cobrados por esse exame (CASARIL, 2016) e na relevância dos dados do Pisa na implementação de políticas educacionais (ADDEY, 2016).

Os estudos evidenciam que os efeitos dos exames estandardizados internacionais podem ser analisados de diferentes maneiras, dada sua complexidade. Assim, os autores que assumem uma análise macro abordam o tema considerando os modelos, as teorias e as concepções que oferecem suporte para essas políticas de avaliação, fazendo uma crítica a ela (AFONSO, 2009; CORREIA; ARELARO; FREITAS, 2015) ou indicando modelos e teorias para aprimorar sua eficiência (RUTKOWSKI; DELANDSHERE, 2016; VAN RIJN; SINHARAY; HABERMAN; JOHNSON, 2016). Além disso, na leitura dos artigos (ARAÚJO; TENÓRIO, 2013; CASARIL, 2016), captou-se a preocupação dos autores sobre o modo como um exame estandardizado de efeito

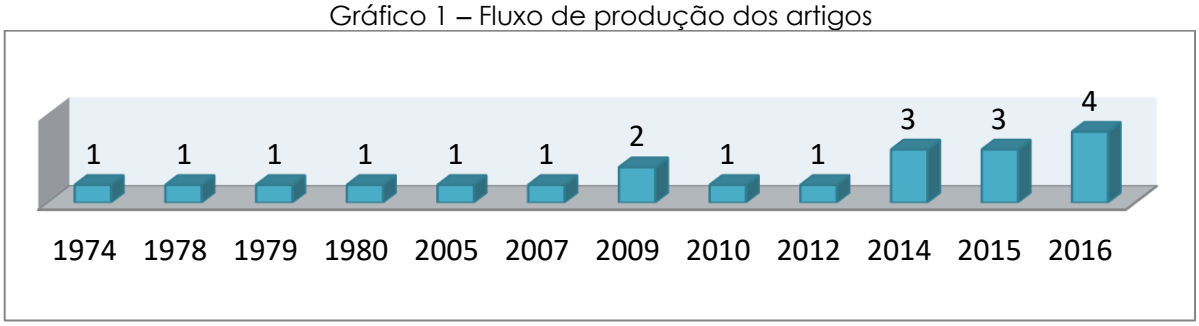
[3] O Pisa é uma iniciativa de avaliação comparada, aplicada em 65 países de forma amostral a estudantes de 15 anos de idade matriculados a partir do sétimo ano do ensino fundamental até o terceiro ano do ensino médio. Tem como objetivo produzir indicadores que contribuam para a discussão da qualidade da educação nos países participantes, de modo a subsidiar políticas de melhoria da educação básica.

[4] O TIMMSS corresponde a uma avaliação internacional que faz a análise comparativa do desempenho dos alunos em Matemática e Ciências em mais de 60 países. O exame recomenda que os países participantes tomem decisões baseadas nos seus resultados para melhorar a política educacional, medindo a eficácia de seus sistemas de ensino em um contexto global, identificando lacunas nos recursos, oportunidade de aprendizagem e as áreas com necessidades para promover as reformas curriculares.

low stakes vem sendo assumido para definir ações que visam a mudanças no âmbito do currículo, da avaliação do ensino e da aprendizagem, sinalizando, dessa maneira, a transição para um efeito do tipo *moderate stakes*.

Para dar visibilidade ao modo como os estudos sobre exames estandardizados vem se constituindo ao longo dos anos elaborou-se o Gráfico 1.

Gráfico 1 – Fluxo de produção dos artigos



Fonte: Os autores (2019).

O Gráfico 1 revela uma produção praticamente linear entre os anos de 1974 e 2007, e 2010 a 2012, com aumento exponencial em 2009 e, gradativamente, em 2014 e 2015, atingindo o pico de produção em 2016. O significativo aumento na última década é revelador de um movimento que vem se ampliando e intensificando em relação à análise dos exames *estandardizados* de natureza nacional e internacional.

Baird, Newton, Hopfenbeck e Tobart (2014), ao estudarem as influências dos testes internacionais e da *Avaliação para a Aprendizagem* (*Assessment for Learning* – AfL)[5], revelam que essas duas perspectivas de avaliação estão conectadas teórica, empírica e conceitualmente com as teorias da aprendizagem. Os autores evidenciam que em diferentes países se têm implementado, cada vez, mais programas nacionais de exames estandardizados. Ao mesmo tempo, houve um aumento dos testes internacionais (Pisa, por exemplo) no início do século XXI, influenciando não só o discurso sobre avaliação, mas impactando, inclusive, os currículos aprovados e a aprendizagem (BAIRD; NEWTON; HOPFENBECK; TOBART, 2014), questão esta também problematizada por Ozga (2012).

Para os autores, os testes internacionais têm deixado cada vez mais evidente que sua finalidade está essencialmente em medir sistemas educacionais e seu impacto sobre a definição do que se compreende por aprendizagem nos níveis

[5] Para os autores, a avaliação para a aprendizagem está no nível de sala de aula, envolvendo uma ampla gama de professores, porém está intimamente ligada à aprendizagem direta do aluno obtida por meio do uso de diferentes instrumentos.

governamentais. Ao mesmo tempo, Baird, Newton, Hopfenbeck e Tobart (2014) destacam que a *Avaliação para a Aprendizagem* tem recebido menos atenção por parte dessas políticas de governo. Para atender aos propósitos dos exames estandardizados (produção de metadados) e melhorar o ranqueamento dos países e suas instituições, esses exames são tomados como referência para a definição da avaliação dos processos de ensino e aprendizagem no contexto escolar em diferentes partes do mundo. Há, nesse caso, adequações das avaliações de ensino e aprendizagem realizadas em sala de aula aos modelos característicos dos exames estandardizados.

Araújo e Fernandes (2009), ao investigarem o processo de implementação e consolidação dos exames estandardizados no Brasil, destacam que a primeira avaliação com essa característica foi criada no país no final da década de 1980, com implementação em 1994, com o Sistema Nacional de Avaliação da Educação Básica (Saeb). Segundo as autoras, esse exame passou por mudanças em sua perspectiva no decorrer dos anos. Nos dois primeiros ciclos do ensino fundamental (primeira, terceira, quinta e sétima séries), a avaliação do desempenho tinha um caráter processual, "[...] já nos dois últimos [oitava e nona séries], a avaliação adquire um caráter 'conclusivo', 'terminativo', indicando a ênfase que, a partir daí e continuamente, seria a ênfase dada aos resultados, bem como ao monitoramento destes" (ARAÚJO; FERNANDES, 2009, p. 128).

Ainda de acordo com as autoras, depois surgiu no Brasil o Exame Nacional de Cursos (ENC – PROVÃO - 1996-2003), denominado, posteriormente, de Exame Nacional de Desempenho dos Estudantes (Enade). Em seguida, devido à ampliação da sua abrangência, ele se integrou ao Sistema Nacional de Avaliação da Educação Superior (Sinaes). Outra avaliação criada no ano de 1998 foi o Exame Nacional do Ensino Médio (Enem), que cambiou de seu caráter terminativo para um sistema de pontuação do vestibular.

Araújo e Fernandes (2009) ressaltam que a crescente utilização desses indicadores, como ferramenta de avaliação da qualidade dos sistemas de ensino, está mais atrelada às políticas educacionais, principalmente quando a finalidade se constitui na comparação desses indicadores. Entretanto, é possível acrescentar que, no caso do Enem, há uma transição de um exame cujo propósito inicial era a produção de metadados, com efeito *low stakes*, para o propósito de certificação

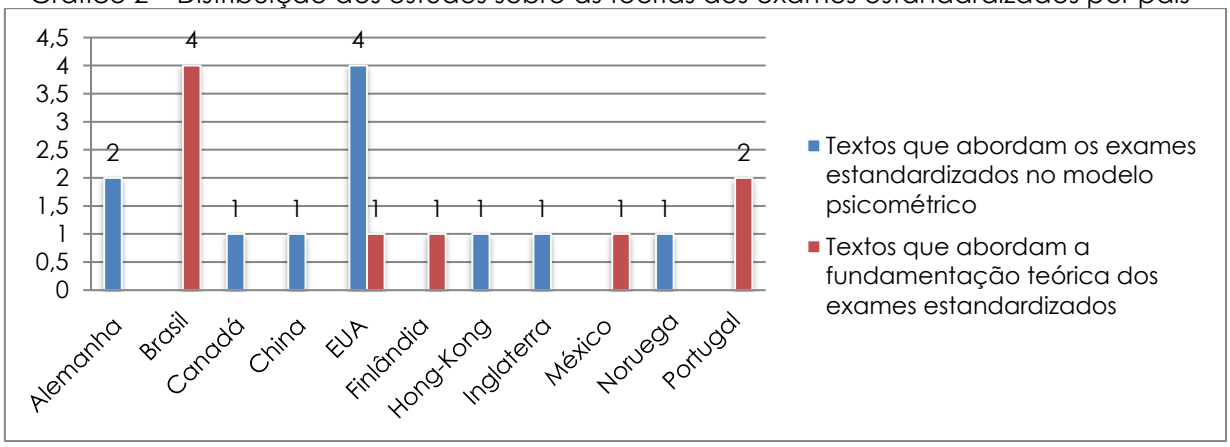
que permite, na maioria dos casos, o acesso dos alunos ao ensino superior, assumindo um efeito do tipo *high stakes*.

De acordo com Retorta (2010), o Saeb e o Sinaes também são considerados exames de grande impacto (*high stakes*), visto que geralmente eles servem como instrumento de classificação e seleção de alunos (Sinaes), ou como instrumento de diagnóstico para detectar problemas e, a partir dos resultados, formular políticas que visam à melhora da qualidade do ensino (Saeb).

De acordo com Fernandes (2007), em Portugal, pelo menos três exames estandardizados internacionais têm sido realizados, incluindo: o TIMSS, a *Second International Assessment of Educational Progress* (Siaep) e o Pisa. Além disso, são aplicados outros exames estandardizados (*high stake*) como: a avaliação de Matemática e de Língua Portuguesa no quarto e no sexto anos de escolaridade; e os exames nacionais no final da escolaridade obrigatória e no final do ensino secundário. Para o autor, essas avaliações estão destinadas à classificação e à certificação dos alunos, ocorrendo geralmente ao final de cada ciclo de escolarização, visando a balanços globais sobre o que os alunos sabem e são capazes de fazer.

A complexidade de realidades e de possibilidades nos estudos sobre exames estandardizados, mesmo após a sua categorização, mostrou que eles apresentam desdobramentos da temática principal a partir das aproximações e dos distanciamentos dos seus conteúdos, o que permitiu agrupar os textos em duas categorias: a) *exames estandardizados no modelo psicométrico* (11); b) *teorias que fundamentam os exames estandardizados* (9).

Gráfico 2 – Distribuição dos estudos sobre as teorias dos exames estandardizados por país



Fonte: Os autores (2019).

Conforme o Gráfico 2, os textos que tematizam os exames estandardizados no modelo psicométrico assumiram, como contexto de estudo, sete países: Estados Unidos da América – EUA (quatro estudos), Alemanha (dois) e Canadá, China, Hong Kong, Inglaterra e Noruega com um estudo cada um. Já os textos que abordam a fundamentação teórica dos exames estandardizados se concentraram em cinco países, dentre eles: o Brasil (quatro), seguido de Portugal (dois) e EUA, Finlândia e México com cada país apresentando um estudo.

Dos 11 países que foram assumidos nos artigos como contextos de estudo do tema, apenas os EUA apresentam estudos das duas categorias de textos. Os demais evidenciam as tendências dos exames estandardizados e seus desdobramentos nos diferentes países, de modo que o modelo psicométrico é o mais recorrente.

Exames estandardizados no modelo psicométrico

Como forma de analisar os exames estandardizados, destacam-se os modelos psicométricos, que se fundamentam na teoria da avaliação por mensuração (SILVA; GOMES, 2018) de atitudes, comportamentos, emoções, representações, opiniões, dentre outros. Objetivam-se, por meio deles, explicar o sentido que têm as respostas dadas pelos sujeitos a uma série de tarefas e propor técnicas de medida dos processos educacionais (PASQUALI, 2009).

No contexto dos EUA, os modelos psicométricos de avalição (WORTHEN; SANDER; FITZPATRICK, 2004; SUPOVITZ; TAYLOR, 2005; KAPLAN, 2016; RUTKOWSKI; DELANDSHERE, 2016; VAN RIJN; SINHARAY; HABERMAN; JOHNSON, 2016) são utilizados para fazer a meta-análise dos dados advindos dos exames estandardizados. A análise dos textos mapeados em um primeiro movimento evidencia que os EUA continuam operando na tradição de testar, estratificar, medir, mensurar e quantificar dados em seus sistemas educacionais. Essa tradição de testar teve início na década de 1920 por intermédio de Thorndike, que desenvolveu a teoria da *Emotional Intelligence* (inteligência emocional), visando a mensurar as mudanças comportamentais. Também Ralph Tyler, na década de 1940, fez uso do conceito, defendendo a inclusão de uma variedade de procedimentos avaliativos, tais como: testes, escalas de atitude, inventários, questionários, fichas, registros e outras formas de coletar evidências sobre o rendimento dos alunos em uma perspectiva longitudinal.

Dentre as nomenclaturas apontados nos textos, estão: *Randomized Controlled Trial* (RCT) (RUTKOWSKI; DELANDSHERE, 2016), utilizado em aplicações regulares, aquelas que não envolvem amostragem complexa ou matricial; *Item Response Theory* (IRT) (KAPLAN, 2016; VAN RIJN; SINHARAY; HABERMAN; JOHNSON, 2016), que analisa os metadados estatisticamente; o *Systematic Programs* (SUPOVITZ; TAYLOR, 2005), que complementa programas e políticas para produzir efeitos reforçadores e sinérgicos; e avaliação de programas (WORTHEN; SANDER; FITZPATRICK, 2004), que indica em que e quando os exames serão aplicados e as formas como serão analisados os metadados.

Todas essas terminologias que se fundamentam no modelo psicométrico são utilizadas para analisar os metadados dos exames estandardizados de natureza nacional, para que sejam tomadas decisões baseadas em evidências contabilizadas e relevantes sobre o contexto, insumo, processo e resultado do sistema de interesse. Fica evidente que, nos EUA, o dilema não é qual teoria será utilizada para fundamentar os exames estandardizados (WORTHEN; SANDER; FITZPATRICK, 2004), mas em que e como as variáveis causais podem ser usadas de modo confiável (KAPLAN, 2016).

Já os dois estudos da Alemanha (DAVE, 1980; WEIRICH; HECHT; HAAG; BÖHME, 2014), de acordo com o Gráfico 2, apontam para a necessidade de comparar os métodos dos programas ou projetos de avaliação, desde que estejam baseados nos princípios de pré-planeamento, planejamento, execução e assimilação. Para os autores, os planejadores, os administradores e os avaliadores têm de assumir uma abordagem mais abrangente no sistema de avaliação orientado para aperfeiçoar a eficiência e eficácia das atividades educacionais.

Nesse caso, faz-se necessário estabelecer a comparação dos diferentes métodos de análises chamados de "imputação" para manipular dados ausentes nas diferentes variáveis, objetivando a máxima utilização possível dos *feedbacks* e mudanças. Esse modo de compreensão é característico do modelo de avaliação analítica definido por Daniel Stufflebeam, na década de 1970, como contexto, *input*, processo e produto – CIPP, cuja centralidade está em descrever, obter e proporcionar informações úteis para assim apontar decisões alternativas.

Como nos EUA, na Alemanha é empregado o IRT, segundo Weirich, Hecht, Haag, Böhme (2014), para analisar os metadados de exames estandardizados como: a Avaliação Nacional de Progresso, o Estudo de Tendências em Matemática e Ciências

Internacionais e o Pisa. Conforme os autores, o método é empregado para medir domínios de leitura e ciências, por exemplo, com a finalidade de monitoramento do sistema educacional. A análise do modelo IRT possibilita estabelecer uma conexão entre os exames de natureza nacional e internacional, já que ele oferece as bases para a sua correção.

Contudo, Dave (1980) explica que os resultados dos exames aplicados ao final do programa, como é o caso do IRT, atualmente utilizado na Alemanha, não ajudam na identificação de barreiras e gargalos que surgem para que sejam aplicadas medidas corretivas que melhorem os resultados dos futuros exames. Para o autor, é necessário rejeitar modelos restritivos e abordar procedimentos mais eficazes e dinâmicos.

Na China (CHEN; LA TORRE, 2014) e em Hong Kong (LAM, 2013), são utilizados os Modelos de Diagnóstico Cognitivo (CDM), que são métodos psicométricos desenvolvidos principalmente para serem utilizados nos exames estandardizados, reunindo um conjunto de habilidades ou atributos dentro de um determinado domínio. De acordo com os autores, o modelo pode ser aplicado para diferentes fins de diagnóstico, facilitando a medição da aprendizagem do aluno e ajudando na concepção de uma melhor instrução por parte do professor.

Na Noruega, Rutkowski e Delashere (2016) apontam a existência de métodos de análises dos metadados: a abordagem das variáveis instrumentais e os escores de propensão. Ambos desenvolvidos por estatísticos, economistas e outros cientistas sociais para tratamento dos dados obtidos nos exames estandardizados em escala nacional e internacional. Diante disso, os autores questionam como seria possível avaliar o valor e as finalidades da educação em sociedades distintas em aspectos sociais, econômicos e culturais.

Fraser (1974) na Inglaterra e Misanchuck (1978) no Canadá discutem os critérios que orientam a seleção de instrumentos e modelo de análise para avaliar os pontos fortes e fracos de um conjunto de variáveis que contemplam as teorias dos exames estandardizados, em específico o modelo CDM e o de Inferências Causais (Ilsa).

Nesse sentido, fica evidente que os modelos psicométricos utilizados nos exames estandardizados, em diferentes países, servem para aferir os metadados, uma abordagem quantitativa por amostragem. Entende-se, assim, como Fernandes (2009), que existe, para além dos modelos psicométricos, quantitativos por

excelência, tradições, representações, desconfianças, expectativas, disponibilidades e etapas diferenciadas de desenvolvimento (social, cultural, político e moral), e nem tudo que conta em educação pode ser comparado ou mensurado.

A diversidade de achados indica que, apesar de um processo globalizado e em ampla expansão, os usos que são feitos dos exames estandardizados, nesses diferentes países, encontram-se em momentos distintos, dadas as necessidades e realidades de cada sistema educativo.

Identifica-se que não existe um único modelo ou método de análise – Inferências Causais (RUTKOWSKI; DELANDSHERE, 2016), Avaliação do Ajuste da Resposta ao Item (VAN RIJN; SINHARAY; HABERMAN; JOHNSON, 2016) ou Imputação Múltipla (WEIRICH; HECHT; HAAG; BÖHME, 2014) – que atenda a todas as variáveis educacionais de um país, por exemplo: escola, currículo, professores e políticas públicas.

Entende-se, ainda, que existe no modelo psicométrico uma articulação entre as teorias da aprendizagem, sobretudo do campo da Psicologia, e os conhecimentos matemáticos. Ou seja, por ser essencialmente um modelo que se fundamenta nas técnicas e na métrica matemática, preocupa-se mais com os resultados em vez de processos de aprendizagem. Portanto, o que está em discussão não é uma crítica ao modelo psicométrico de avaliação, mas a necessidade de saber como esses dados são interpretados e quais políticas são desenvolvidas a partir deles.

Teorias que fundamentam os exames estandardizados

Para estudiosos do campo da avaliação, os exames estandardizados são influenciados por diferentes teorias educacionais, visando à elaboração de modelos para valorar políticas, programas, projetos e ações educativas. Desse modo, os estudos do Brasil (BAUER; ALAVARSE; OLIVEIRA, 2015), EUA (EBEL, 1979), Portugal (AFONSO, 2009; CORREIA; ARELARO; FREITAS, 2015), México (OLIVÓS, 2014) e Finlândia (SALOKANGAS; KAUKO, 2015) têm dado visibilidade às teorias que vêm fundamentando os exames estandardizados nesses contextos.

Partindo do questionamento: o que justifica a adoção de diferentes teorias avaliativas em países distintos em que há exames estandardizados nacionais e/ou internacionais? Acredita-se que uma teoria que seja baseada nas Representações Sociais (FERREIRA; TENÓRIO, 2010), na avaliação do desempenho (TOFFOLI; ANDRADE; BORNIA; QUEVEDO-CAMARGO, 2016), no modelo *accountability*

(AFONSO, 2009) ou na teoria do paradigma científico[6] (CORREIA; ARELARO; FREITAS, 2015), dificilmente irá atender às diferentes demandas e particularidades culturais, econômicas ou políticas desses países.

Especificamente no caso brasileiro, quatro estudos (FONTANIVE; ELLIOT; KLEIN, 2007; FERREIRA; TENÓRIO, 2010; BAUER; ALAVERSE; OLIVEIRA, 2015; TOFFOLI; ANDRADE; BORNIA; QUEVEDO-CAMARGO, 2016) focalizam a visão epistemológica na construção de políticas de avaliação, na sistematização do debate e em programas específicos de avaliação. O produto ainda é permeado de muitas dúvidas e incertezas: como é realizado? Quais são os métodos de análises dos metadados utilizados? Qual a forma de contratação dos avaliadores e a devolutiva desses resultados para a formação de novas políticas educacionais e, consequentemente, de novos currículos?

Os dados revelam que, nos EUA, já não é o processo, e sim o produto (dados psicométricos) da construção e fundamentação teórica dos exames estandardizados a grande problemática a ser investigada, uma vez que apenas um estudo se propõe fazer esse tipo de análise nesse contexto (EBEL, 1979). Porém, Souza e Vasquez (2015) salientam que os modelos mais eficientes resultantes da meta-análise visam a atender e responder, com maior precisão, às inúmeras variáveis socioemocionais: (a ansiedade e autoeficácia) e as sociodemográficas e do contexto escolar (qualidade da instrução e a percepção de prazer, raiva, tédio e ansiedade) que influenciam o resultado final dos exames.

Aprofundando-se na discussão sobre as teorias avaliativas e, consequentemente, sobre os modelos que fundamentam os exames estandardizados, é preciso esclarecer que as discussões dos textos revelam preocupações: com a construção de indicadores que tenham como referências bases teóricas consistentes e as formas de avaliá-las (FERREIRA; TENÓRIO, 2010); com a construção de políticas intermediadas por imposição de modelos avaliativos nacionais e internacionais (BAUER; ALAVARSE; OLIVEIRA, 2015), e com as contribuições e críticas que permeiam esses diferentes modelos (EBEL, 1979; SALOKANGAS; KAUKO, 2015; TOFFOLI; ANDRADE; BORNIA; QUEVEDO-CAMARGO, 2016).

[6] Os exames estandardizados, cada dia mais, vêm ganhando *status* de critério único e científico (por que avaliar e como) para o que crianças e jovens aprendem na escola (CORREIA; ARELARO; FREITAS, 2015).

Nesse sentido, é evidenciado, por Ferreira e Tenório (2010), que a complexidade se dá, também, em decorrência da multidimensionalidade do real social, cujos recortes analíticos são sempre provisórios, situados, restritivos, permitindo apenas uma avaliação parcial dos fenômenos.

Assim, a teoria baseada nas Representações Sociais, segundo Ferreira e Tenório (2010), aborda a construção de um modelo voltado para valorar políticas educativas, fundamentado em aspectos vinculados aos interesses sociopolíticos de determinados grupos em confronto e suas representações sobre a qualidade da educação. Desse modo, os exames estandardizados têm como objetivo construir instrumentos e indicadores de qualidade que possam exprimir os aspectos objetivos da realidade, mas também apreender as representações e interesses em jogo, visando a lograr os aspectos qualitativos e quantitativos.

> Um Indicador se revela, portanto, como um elemento, sinal ou aviso que revela ou denota características especiais ou qualidades, que apontam (como o *dedo indicador*) uma direção, mostrando a conveniência de, ou aconselhando a alguma ação. De forma mais técnica, é um composto construído para medir uma dimensão ou variável (FERREIRA; TENÓRIO, 2010, p. 73).

Os autores evidenciam que não é só difícil avaliar o valor e as finalidades da educação somente em sociedades distintas, mas também internamente, pois não podem ser desconsiderados os embates promovidos pelas relações de forças (GINZBURG, 2002) em busca das articulações, representações e hegemonias do domínio sobre o lugar[7].

Já a avaliação do desempenho, baseada na Teoria Utilitária, segundo Toffoli, Andrade, Bornia e Quevedo-Camargo (2016), é assim designada por ter, como base principal, a necessidade de especialistas para corrigir as respostas. É utilizada em uma variedade de áreas, por exemplo, em competições esportivas, cujos critérios são preestabelecidos nos testes com itens de respostas abertas. Enquadram-se, também, na correção das redações em vestibulares e de outros concursos, testes orais e entrevistas para seleção. Conforme os autores, esses exames possuem diferentes objetivos, dentre eles, determinar o grau de habilidade para uma atividade

[7] "Um lugar é a ordem (seja qual for), segundo a qual se distribuem elementos nas relações de coexistência. Aí se acha, portanto, excluída a possibilidade, para duas coisas, de ocuparem o mesmo lugar. Aí impera a lei do 'próprio': os elementos considerados se acham uns ao lado dos outros, cada um situado num lugar 'próprio' e distinto que define. Um lugar é, portanto, uma configuração instantânea de posições. Implica uma indicação de estabilidade" (CERTEAU, 1994, p. 201).

específica, portanto é preciso que as informações provenientes desses testes sejam confiáveis, pois eles auxiliam as decisões de pessoas ou da esfera pública.

O modelo avaliativo denominado de *Accountability*, fundamentado na Teoria da Responsabilização, é o mais mencionado na literatura (PAULSON; MARCHANT, 2009; FERREIRA; TENÓRIO, 2010; SALOKANGAS; KAUKO, 2015), porém é no estudo de Afonso (2009) que a discussão sobre o modelo, baseado em testes estandardizados e rankings escolares, se encontra estruturada[8].

De acordo com Afonso (2009), o conceito de *Accountability* é em geral polissêmico e denso, associado a três dimensões articuláveis: a) avaliação; b) prestação de contas; e c) responsabilização. Como é o principal modelo de avaliação do sistema educativo norte-americano e em implantação no contexto português, Afonso (2009) defende uma linha de reflexão e pesquisa que se assente em uma concepção *Accountability* mais ampla, fundamentada e complexa do ponto de vista teórico-metodológico, político, axiológico e epistemológico, de modo que se torne relativamente imune à ideologia política, como ocorreu no caso da Inglaterra, em que tanto os governos do *New Labour*[9] como os governos do Partido Conservador lhe deram grande ênfase.

Acreditando que nem tudo que conta em educação é mensurável ou comparável, Afonso (2009) critica os exames estandardizados baseados no modelo *Accountability* para a criação de *rankings*. A crítica que o autor faz a esse modelo é de que os resultados se dão de forma parcelar, incompleta e redutora, em face à complexidade e pluralidade dos objetivos, missões e funções da educação escolar.

Percebe-se que, na Teoria da Responsabilização, a presença do Estado-avaliador que assume explicitamente o exercício e controle social como regulador do sistema educativo, visando à construção de políticas públicas, com mais força do que nas teorias Representações Sociais e Utilitárias. Olivós (2014), analisando os impactos dos exames estandardizados em países como o México, os EUA e a Escócia, e os pesquisadores Salokangas e Kauko (2015), investigando trabalhos sobre o tema

[8] Segundo Becker (2010), o modelo *Accountability* na educação foi pioneiro, criado em 1988, na Inglaterra, por meio do *Educational Reform Act.* Nessa reforma, houve a centralização do currículo, criação de sistemas de avaliação e, aos poucos, as escolas passaram a ter mais liberdade para gerir os recursos recebidos.

[9] Nome dado ao Partido Trabalhista do Reino Unido, quando liderado por Tony Blair, para mostrar uma renovação do partido com uma tendência menos de esquerda e mais moderna (Fonte: http://pt.bab.la/dicionario/ingles-portugues/new-labour).

na Finlândia, revelam que as escolas nesses contextos, para alcançar objetivos, controlam e prescrevem, com uma abordagem mais aberta, o modelo de responsabilização, acompanhado de requisitos burocráticos. Consequentemente, os desdobramentos dos exames estandardizados assumem diferentes critérios e modos de avaliar o processo de ensino interno, externo, formativo e somativo, cujos resultados impactam diretamente os sistemas educacionais desses países.

Diante desse contexto, é preciso considerar que, seja qual for a teoria utilizada, ela apresentará fragilidades no momento da meta-análise, por demandar inúmeras variáveis que estabelecem relação com o produto e também com o processo, o que exige diferentes ferramentas e métodos de análises.

Nesse sentido, para Correia, Arelaro e Freitas (2015), os exames estandardizados são processos complexos e, para serem bem-sucedidos, necessitam da participação consciente dos envolvidos e ética nos processos avaliativos, já que, por suas exigências incompatíveis com o cotidiano escolar, têm levado à "corrupção" na educação e à "distorção" na análise de programas como o Pisa. Destacam ainda que o Pisa, fundamentado no modelo psicométrico, vem influenciando a transformação e os modos de pensar as avaliações e as políticas públicas de educação, dando atenção especial aos sujeitos envolvidos, professores e estudantes.

De maneira geral, os textos dessa categoria estão dedicados à discussão das concepções, às intencionalidades e aos usos que tem sido feito dos exames estandardizados. Para Baird, Newton, Hopfenbeck e Tobart (2014), muitas práticas avaliativas não têm sido orientadas pela teoria, mesmo que possa ter havido suposições implícitas sobre aprendizagem e o papel da avaliação nesse processo.

Compreende-se que os estudos, ao discutirem as teorias que fundamentam os exames estandardizados, geralmente estabelecem críticas. A primeira delas corresponde à falta de clareza teórica desses exames que têm seus desdobramentos em determinadas concepções, o que, consequentemente, vai direcionar o que e para que essas avaliações são realizadas. A segunda crítica está no fato de que as intencionalidades desses exames estão fortemente alinhadas à ideia da mensuração. Além disso, identificam que ainda não se tem definido como se operam os dados originários desses exames, sobretudo no que se refere ao que se busca identificar e que ações ou políticas educacionais são encaminhadas a partir da análise desses resultados.

Conclusão

Este estudo teve por objetivo analisar como tem se constituído o debate dos exames estandardizados em diferentes países. Com isso, identificou-se que as pesquisas sobre o tema vão apresentando uma diversidade de modelos que tem relação com questões históricas e políticas específicas de cada país, tendo como reflexos a criação de diferentes teorias de avaliação.

Independentemente da nomenclatura utilizada para identificá-los, nos diferentes países, ficou evidente que esses exames são geradores da discussão sobre os aspectos referentes às peculiaridades dos sistemas educacionais, como: a) teorias e sistemas que fundamentam os processos avaliativos; b) métodos de análise dos metadados, seus usos e aplicações; c) variáveis sociodemográfica, socioemocional e fracasso escolar. Essas são derivações abordadas pelos autores, evidenciando a complexidade desse fenômeno que vem ganhando *status* globalizado, sem levar em consideração os diferentes dilemas dos países em frente aos desafios associados a tais exames.

Na Europa e nos EUA, por exemplo, a discussão sobre o fenômeno "exames estandardizados" está estruturada nos modelos psicométricos (ênfase no produto). Estudam-se métodos de análises dos metadados, visando a contemplar, com rigor e precisão, as diferentes variáveis intervenientes que estabelecem relação direta com os testes e servem de base para o aprimoramento das diretrizes que fazem valer as políticas públicas de avaliação.

Já na América Latina, em países como Chile, México e, principalmente, Brasil, a discussão sobre os exames estandardizados está estruturada com ênfase nos aspectos históricos, nas diferentes políticas e programas de avaliação, nos currículos e em seus desdobramentos para os sistemas educativos e os sujeitos que os compõem, ou seja, no processo, sendo o produto pouco compreendido, inexplorado e não transparente.

Há uma tendência "perigosa" que leva os sistemas de ensino para um vicioso círculo, pautado na elaboração dos currículos com base naquilo que os exames estandardizados apontam como deficiência dos alunos. Assim, para que não haja resultados negativos nos testes, políticas públicas têm direcionado a prática dos professores fundamentada no que será cobrado nos exames. Ou seja, os professores passam a ter pouca autonomia para escolher os conteúdos que consideram

importantes para atender às individualidades dos seus alunos, e estes, ao estudarem para os exames, não atribuem novos sentidos àquilo que aprendem.

Nesse caso, entende-se que essas discrepâncias encontradas entre a história de uma política pública de avaliação e outra que se encontra em vigor revelam fragilidades e se tornam suscetíveis a duras críticas. Geralmente as análises e os resultados dos exames estandardizados são puramente baseados em métodos quantitativos, o que justifica a complementação pelo método qualitativo dos metadados (ALTMANN, 2002; ESTEBAN, 2009; COSGROVE; CARTERIGHT, 2014). Por fim, faz-se necessária uma maior exploração por parte dos professores sobre o produto final dos exames, no sentido de aperfeiçoar os usos que deles possam ser extraídos.

Referências

ADDEY, C. O Pisa par ao desenvolvimento e o sacrifício de dados com relevância política. *Educação e Sociedade*, Campinas, SP, v. 37, n. 136, p. 685-706, jul./set. 2016. DOI: https://doi.org/10.1590/es0101-73302016166001. Disponível em: http://www.scielo.br/scielo.php?pid=S0101-73302016000300685&script=sci_abstract&tlng=pt. Acesso em: 23 fev. 2017.

AFONSO, A. J. Nem tudo o que conta em educação é mensurável ou comparável: crítica à accountability baseada em testes estandardizados e rankings escolares. *Revista Lusófona de Educação*, Lisboa, v. 13, n. 13, p. 13-29, 2009. Disponível em: https://revistas.ulusofona.pt/index.php/rleducacao/issue/view/57. Acesso em: 23 fev. 2017.

ALTMANN, H. Influências do Banco Mundial no projeto educacional brasileiro. *Educação e Pesquisa*, São Paulo, v. 28, n. 1, p. 77-89, jan./jun. 2002. Disponível em: http://www.scielo.br/pdf/ep/v28n1/11656.pdf. Acesso em: 23 fev. 2017.

ARANCIBIA, V. *Los sistemas de medición y evaluación de calidad de la educación*. Santiago: OREALC, 1997.

ARAÚJO, G. C.; FERNADES, C. F. R. Qualidade do ensino e avaliações em larga escala no Brasil: os desafios do processo e do sucesso educativo na garantia do direito à educação. *Revista Iberoamericana de Evaluación Educativa*, Madrid, v. 2, n. 2, p. 125-140, out. 2009. Disponível em: https://revistas.uam.es/index.php/riee/article/view/4562. Acesso em: 2 fev. 2017.

ARAÚJO, M. de L. H. S.; TENÓRIO, R. M. Resultados brasileiros no Pisa e seus (des)usos. *Estudos em Avaliação Educacional*, São Paulo, v. 28, n.68, p. 344-380, maio/ago. 2009. Disponível em: http://publicacoes.fcc.org.br/ojs/index.php/eae/article/view/4553. Acesso em: 2 fev. 2017.

AUGUSTO, M. H. Regulação educativa e trabalho docente em Minas Gerais: a obrigação de resultados. *Educação e Pesquisa*, São Paulo, v. 38, n. 03, p. 695-709, jul./set. 2012.

BARROS, M. C. M. M. de; TAVARES, P. de A.; MASSEI, W. O desenvolvimento da educação no estado de São Paulo: sistema de avaliação do rendimento escolar, plano de desenvolvimento da educação e bonificação variável por desempenho. *São Paulo em Perspectiva*, São Paulo, v. 23, n. 1, p. 42-56, jan./jun. 2009. Disponível em: http://produtos.seade.gov.br/produtos/spp/v23n01/v23n01_04.pdf. Acesso em: 3 mar. 2017.

BAUER, A.; ALAVARSE, O. M.; OLIVEIRA, R. P. Avaliação em larga escala: uma sistematização do debate. *Educação e Pesquisa*, São Paulo, v. 41, n. especial, p. 1367-1382, dez. 2015. Disponível em: http://www.scielo.br/pdf/ep/v41nspe/1517-9702-ep-41-spe-1367.pdf. Acesso em: 3 mar. 2017.

BECKER, F. R. Avaliação educacional em larga escala: a experiência brasileira. *Revista Iberoamericana de Educação*, Madrid, v. 53, n. 1, p. 1-11, jun. 2010. DOI: https://doi.org/10.35362/rie5311751. Disponível em: https://rieoei.org/RIE/issue/view/145. Acesso em: 3 mar. 2017.

BECKER, F. R. Avaliações externas e ensino fundamental: do currículo para a qualidade ou da "qualidade" para ao currículo. REICE. *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*, Madrid, v. 10, n. 4, p. 37-48, out. 2012. Disponível em: https://revistas.uam.es/index.php/reice/article/view/2986. Acesso em: 3 mar. 2017.

BAIRD, J.; NEWTON, P. E.; HOPFENBECK, T. N.; STOBART, G. *Assessment and learning*: state of the field review. [Oxford]: Knowledge Center for Education, 2014.

BONAMINO, A.; SOUZA, S. Z. Três gerações de avaliação da educação básica no Brasil: interfaces com o currículo da/na escola. *Educação e Pesquisa*, São Paulo, v. 38, n. 2, p. 373-388, abr./jun. 2012. Disponível em: http://www.scielo.br/pdf/ep/v38n2/aopep633.pdf. Acesso em: 3 mar. 2017.

BROOKE, N. Responsabilização educacional no Brasil. *Revista Iberoamericana de Evaluación Educativa*, Madrid, v. 1, n. 1, p. 94-109, 2008. Disponível em: https://revistas.uam.es/index.php/riee/article/view/4684/5120. Acesso em: 13 ago. 2016.

CAICEO ESCUDERO, J. Los sistemas estandardizados de evaluación en Chile: participación de Mario Leyton Soto y Erika Himmel König. Historia de la Educación: *Revista Interuniversitaria*, Salamanca, n. 34, p. 357-371. 2015. DOI: http://dx.doi.org/10.14201/hedu201534357371. Disponível em: https://revistas.usal.es/index.php/0212-0267/article/viewFile/hedu201534357371/15712. Acesso em: 3 mar. 2017.

CASTRO, M. H. G. de. Sistemas nacionais de avaliação e de informações educacionais. *São Paulo em Perspectiva*, São Paulo, v. 14, n. 1, p. 121-128, jan./mar. 2000. Disponível em: http://www.scielo.br/pdf/spp/v14n1/9809.pdf. Acesso em: 13 dez. 2018.

CASARIL, M. Programa internacional de avaliação de estudantes (Pisa): a concepção do letramento e o estado da arte no Brasil. *Revista Trama*, Cascável, PR, v. 12, n. 27, p. 84-109, 2016. Disponível em: http://e-revista.unioeste.br/index.php/trama/article/view/14458. Acesso em: 13 dez. 2018.

CATANI, D. B. A imprensa periódica educacional: as revistas de ensino e o estudo do campo educacional. *Educação e Filosofia*, Uberlândia, MG, p.115-130, jul./dez. 1996. Disponível em: http://www.seer.ufu.br/index.php/EducacaoFilosofia/issue/view/83. Acesso em: 2 fev. 2017.

CERTEAU, M. *A invenção do cotidiano*: volume 1: artes de fazer. 15. ed. Petrópolis, RJ: Vozes, 1994.

CHEN, J. S.; LA TORRE, J. A procedure for diagnostically modeling extant large-scale assessment data: the case of the programme for international student assessment in reading. *Psicology*, Washington, n. 5, p. 1967-1978, nov. 2014.

CORREIA, J. A. de A. e V.; ARELARO, L. R. G.; FREITAS, L. C. de. Para onde caminham as atuais avaliações educacionais?. *Educação e Pesquisa*, São Paulo, v. 41, n. especial, p. 1275-1281, dez. 2015. Disponível em: http://www.scielo.br/pdf/ep/v41nspe/1517-9702-ep-41-spe-1275.pdf. Acesso em: 23 fev. 2017.

COSGROVE, J.; CARTERIGHT, F. Changes in achievement on PISA: the case of Ireland and implications for international assessment practice. *Large-scale Assessments in Education*, *[S. l.]*, v. 2, n. 2, p. 2-17, jan. 2014.

DAVE, R. H. A built-in system of evaluation for reform projects and programmes in education. *International Review of Education*, Hamburg, v. 26, n. 4, p. 475-482, dec. 1980.

EBEL, R. L. The role of testing in basic education. *NASSP Bulletin*, Nevada, v. 63, n. 429, p. 89-93, 1979.

ESTEBAN, M. T. Avaliação e fracasso escolar: questões para debate sobre a democratização da escola. *Revista Lusófona de Educação*, Lisboa, n. 13, p. 123-134, jun. 2009. Disponível em: http://www.scielo.mec.pt/pdf/rle/n13/13a08.pdf. Acesso em: 23 mar. 2017.

FERNANDES, D. A avaliação das aprendizagens no sistema educativo português. *Educação e Pesquisa*, São Paulo, v. 33, n. 3, p. 581-600, set./dez. 2007. Disponível em: http://www.scielo.br/pdf/ep/v33n3/a13v33n3.pdf. Acesso em: 23 fev. 2017.

FERNANDES, D. *Avaliar para aprender*: fundamentos, práticas e políticas. São Paulo: Ed. UNESP, 2009.

FERREIRA, C. A. A avaliação das aprendizagens no ensino básico português e o reforço da avaliação sumativa externa. *Educação e Pesquisa*, São Paulo, v. 41, n. 1, p. 153-169, jan./mar. 2015. DOI: https://doi.org/10.1590/S1517-970220150118 92. Disponível em: http://www.scielo.br/pdf/ep/v41n1/1517-9702-ep-41-1-0153.pdf. Acesso em: 23 mar. 2017.

FERREIRA, R. A.; TENÓRIO, R. M. A construção de indicadores de qualidade no campo da avaliação educacional: um enfoque epistemológico. *Revista Lusófona de Educação*, Lisboa, v. 15, n. 15, p. 71-97, jan. 2010. Disponível em: http://www.scielo.mec.pt/pdf/rle/n15/n15a06.pdf. Acesso em: 23 mar. 2017.

FINI, M. I. Currículo e avaliação: articulação necessária em favor da aprendizagem dos alunos da rede pública de São Paulo. *São Paulo em Perspectiva*, São Paulo, v. 23, n. 1, p. 57-72, jan./jun. 2009. Disponível em: http://produtos.seade.gov.br/produtos/spp/index.php?men=rev&cod=5080. Acesso em: 23 mar. 2017.

FONTANIVE, N. S.; ELLIOT, L. G.; KLEIN, R. Os desafios da apresentação dos resultados da avaliação de sistemas escolares a diferentes públicos. *REICE*: Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, Madrid, v. 5, n. 2, p. 260-273, 2007.

FRASER, B. J. A seleção de instrumentos de avaliação. *Journal Research in Science Education, [S. l.]*, v. 4, n. 1, p. 99-111, 1974.

GINZBURG, C. *Relações de força*: história, retórica e prova. São Paulo: Companhia das Letras, 2002.

GIPPS, C. A avaliação de sistemas educacionais: a experiência inglesa. *In*: CONHOLATO, M. C. (coord.). *Sistemas de avaliação educacional*, São Paulo: FDE, 1998. p. 123-135. Disponível em: http://www.crmariocovas.sp.gov.br/pdf/ideias_30_p123-135_c.pdf. Acesso em: 8 mar. 2018.

KAPLAN, D. Causal inference with large-scale assessments in education from a Bayesian perspective: a review and synthesis. *Large-scale Assessments in Education*, [S. l.], v. 4, n. 7, p. 2-24, may. 2016.

LAM, R. Formative use of summative tests: using test preparation to promote performance and self-regulation. *The Asia-Pacific Education Researcher*, Switzerland, v. 22, n. 1, p. 69-78, feb. 2013.

LAVASSEUR, J. Plano de avaliação do conhecimento dos alunos na França. In: ALMEIDA, Fernando José de (org.). *Avaliação educacional em debate*: experiências no Brasil e na França. São Paulo: Cortez: EDUC, 2005.

MISANCHUCK, E. R. Descriptors of evaluations in instructional development: beyond the formative-summative distinction. *Journal of Instructional Development*, Illinois, v. 2, n. 1, p. 15-19, 1978.

MOROSINI, M. C.; FERNANDES, C. M. B. Estado do conhecimento: conceitos, finalidades e interlocuções. *Educação Por Escrito*, Porto Alegre, v. 5, n. 2, p. 154-164, jul./dez. 2014. Disponível em: http://revistaseletronicas.pucrs.br/ojs/index.php/porescrito/article/view/18875. Acesso em: 12 mar. 2017.

NAVARRO, F.; REYES, N. Á.; VERA, G. G. Validez y justicia: hacia una evaluación significativa en pruebas estandarizadas de escritura. *Revista Meta*: Avaliação, Rio de Janeiro, v. 11, n. 31, p. 1-35, jan./abr. 2019. DOI: http://dx.doi.org/10.22347/2175-2753v11i31.2045. Disponível em: http://revistas.cesgranrio.org.br/index.php/metaavaliacao/article/view/2045. Acesso em: 7 abr. 2019.

OLIVÓS, T. M. Posturas epistemológicas frente a la evaluación y sus implicaciones en el currículum. *Perspectiva Educacional*, Valparaíso, v. 53, n. 1, p. 3-18, jan. 2014. Disponível em:

http://www.perspectivaeducacional.cl/index.php/peducacional/article/viewFile/211/96. Acesso em: 7 mar. 2017.

ORFIELD, G.; WALD, J. Testing, testing: the high-stakes testing mania hurts poor and minority students the most. The Nation, New York, v. 270, n. 22, p. 38-40, 2000.
OZGA, J. Assessing Pisa. *European Educational Research Journal*, Oxford, v. 11, n. 2, p. 166-171, jun. 2012.

PASQUALI, L. Psicometria. *Revista da Escola de Enfermagem*, São Paulo, v. 43, n. especial, p. 992-999, dez. 2009. Disponível em: http://www.scielo.br/pdf/reeusp/v43nspe/a02v43ns.pdf. Acesso em: 11 mar. 2017.

PAULSON, S.; MARCHANT, G. Variáveis de Background, níveis de agregação e pontuações de testes padronizados. *Education Policy Analysis Archives*, Arizona, v. 17, n. 22, p. 1-24, nov. 2009.

RETORTA, M. S. Percepções do professor sobre o SAEB: um estudo sobre o efeito retroativo. *Revista Educação & Tecnologia*, Curitiba, v. 1, p. 133-174, 2010. Disponível em: http://revistas.utfpr.edu.br/pb/index.php/revedutec-ct/article/view/1357. Acesso em: 2 fev. 2020.

RUTKOWSKI, D.; DELANDSHERE, G. Causal inferences with large scale assessment data: using a validity framework. *Large-scale Assessments in Education*, [S. l.], v. 4, n. 6, p. 1-18, may 2016.

SALOKANGAS, M.; KAUKO, J. Tomar de empréstimo o sucesso finlandês no Pisa?: algumas reflexões críticas, da perspectiva de quem faz este empréstimo. *Educação e Pesquisa*, São Paulo, v. 41, n. especial, p. 1353-1365, dez. 2015. DOI: https://doi.org/10.1590/S1517-9702201508144214. Disponível em: http://www.scielo.br/pdf/ep/v41nspe/1517-9702-ep-41-spe-1353.pdf. Acesso em: 23 maio 2017.

SANTOS, A. P. dos. Abordagem do ciclo de políticas e suas contribuições para análise da política de avaliação em larga escala. *Meta*: Avaliação, Rio de Janeiro, v. 6, n. 18, p. 263-280, set./dez. 2014. DOI: https://doi.org/10.1590/S0101-73302006000100003. Disponível em: http://www.scielo.br/pdf/es/v27n94/a03v27n94.pdf. Acesso em: 7 abr. 2019.

SCHNEIDER, M. P. Políticas de avaliação em larga escala e a construção de um currículo nacional para a educação básica. *Eccos Revista Científica*, São Paulo, n. 30, p. 17-33, jan./abr. 2013. DOI: 10.5585/EccoS.n30.3537. Disponível em: https://www.redalyc.org/pdf/715/71525769002.pdf. Acesso em: 23 maio 2017.

SILVA, A. L.; GOMES, A. M. Avaliação educacional: concepções e embates teóricos. *Estudos em Avaliação Educacional*, São Paulo, v. 29, n. 71, p. 350-84, maio/ago. 2018. Disponível em: http://publicacoes.fcc.org.br/ojs/index.php/eae/article/view/5048. Acesso em: 15 dez. 2018.

SILVA, I. F. O sistema nacional de avaliação: características, dispositivos legais e resultados. *Estudos em Avaliação Educacional*, São Paulo, v. 21, n. 47, p. 427-448,

set./dez. 2010. Disponível em: https://www.fcc.org.br/pesquisa/publicacoes/eae/arquivos/1602/1602.pdf. Acesso em: 15 dez. 2018.

SOUZA, D. C. C.; VASQUEZ, D. A. Expectativas de jovens do ensino médio público em relação ao estudo e ao trabalho. *Educação e Pesquisa*, São Paulo, v. 41, n. 2, p. 409-426, abr./jun. 2015. DOI: https://doi.org/10.1590/s1517-97022015041789. Disponível em: http://www.scielo.br/pdf/ep/v41n2/1517-9702-ep-41-2-0409.pdf. Acesso em: 22 mar. 2018.

SUDBRACK, E. M.; COCCO, E. M. Avaliação em larga escala no Brasil: potencial indutor de qualidade? *Roteiro*, Joaçaba, SC, v. 39, n. 2, p. 347-370, jul./dez, 2014. Disponível em: https://pt.scribd.com/document/354104136/AVALIACAO-EM-LARGA-ESCALA-NO-BRASIL-POTENCIAL-INDUTOR-DE-QUALIDADE. Acesso em: 11 set. 2017.

SUPOVITZ, J. A.; TAYLOR, B. S. Systemic education evaluation evaluating the impact of systemwide reform in education. *American Evaluation Association*, Washington, v. 26, n. 2, p. 204-230, jun. 2005.

THORNDIKE, E. L. *The new methods in Arithmetic*. San Francisco: Rand Mcnally & Company, 1921.

TOFFOLI, S. F. L.; ANDRADE, D. F. de; BORNIA, A. C.; QUEVEDO-CAMARGO, G. Avaliação com itens abertos: validade, confiabilidade, comparabilidade e justiça. *Educação e Pesquisa*, São Paulo, v. 42, n. 2, p. 343-358, abr./jun. 2016. DOI: https://doi.org/10.1590/S1517-9702201606135887. Disponível em: http://www.scielo.br/pdf/ep/v42n2/1517-9702-ep-42-2-0343.pdf. Acesso em: 23 mar. 2017.

TORANZOS, L. V. Evaluación educativa: hacia la construcción de un espacio de aprendizaje. *Propuesta Educativa*, Tucumán, v. 23, n. 41, p. 9-19, jun. 2014.

TYLER, R. W. *Basic principles of curriculum and insmction*. Chicago: University of Chicago Press, 1949.

VAN RIJN, P. W.; SINHARAY, S.; HABERMAN, S.; JOHNSON, M. Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-scale Assessments in Education, [S. l.],* v. 4, n. 10, p. 1-23, oct. 2016.

WEIRICH, S.; HECHT, M.; HAAG, N.; BÖHME, K. Nested multiple imputation in large-scale assessments. *Large-scale Assessments in Education, [S. l.]*, v. 2, n. 9, p. 1-18, sep. 2014.

WERLE, F. O. C.; THUM, A. B.; ANDRADE, A. C. de. Processo nacional de avaliação do rendimento escolar: tema esquecido entre os sistemas municipais de ensino. *Ensaio*: Avaliação e Políticas Públicas em Educação, Rio de Janeiro, v. 17, n. 64, p. 397-420, jul./set. 2009. Disponível em: http://www.scielo.br/pdf/ensaio/v17n64/v17n64a02.pdf. Acesso em: 22 mar. 2017.

WORTHEN, B. R.; SANDERS, J. R.; FITZPATRICK, J. L. *Avaliação de programas*: concepções e práticas. São Paulo: Ed. Gente, 2004.