AVALIAÇÃO EM LARGA ESCALA NA AMÉRICA LATINA
LARGE-SCALE EVALUATION IN LATIN AMERICA
EVALUACIÓN EN LARGA ESCALA EN LATINOAMÉRICA

https://doi.org/10.18222/eae.v35.11050_port

UMA BREVE HISTÓRIA DOS TESTES DE ALTO IMPACTO E SEUS POSSÍVEIS FUTUROS

MARÍA JESÚS GUTIÉRREZ DOMÍNGUEZ[I]

Tradução de: Laura Mendes Loureiro

[I] Universitat Autònoma de Barcelona (UAB), Cerdanyola del Vallès, Espanha; *mjgutierrez@uc.cl*
[II] Viamundi Idiomas e Traduções, Belo Horizonte-MG, Brasil; *laura@viamundi.com.br*

RESUMO

A ênfase histórica da educação em avaliações que promovem a aprendizagem evoluiu para um modelo competitivo e orientado pelo desempenho impulsionado por reformas educacionais e influências neoliberais. Essa tendência materializou-se de diversas formas, sendo uma delas os testes de alto impacto (*high-stakes testing*). Por meio de uma Revisão Sistemática da Literatura (RSL), o presente estudo desenvolve uma análise global desse mecanismo, buscando compreender seu alcance e impacto no sistema educacional. Os resultados revelam tanto os efeitos adversos quanto os benefícios dos testes de alto impacto. Embora a literatura destaque consequências não intencionais, também enfatiza a importância dos testes como um mecanismo de responsabilização na sociedade contemporânea, indicando aspectos positivos. Além disso, os achados sugerem que alternativas viáveis para avaliações em larga escala vão além dos cenários de alto impacto.

PALAVRAS-CHAVE AVALIAÇÃO EM LARGA ESCALA • TESTES DE ALTO IMPACTO • AVALIAÇÃO ALTERNATIVA • REVISÃO DE LITERATURA.

COMO CITAR:

Gutiérrez Domínguez, M. J. (2024). Uma breve história dos testes de alto impacto e seus possíveis futuros. *Estudos em Avaliação Educacional*, *35*, Artigo e11050. https://doi.org/10.18222/eae.v35.11050_port

A BRIEF HISTORY OF HIGH-STAKES TESTING AND ITS POSSIBLE FUTURES

ABSTRACT

Education's historical emphasis on assessments that enhance learning has evolved into a competitive, performance-driven model due to educational reforms and neoliberal influences. This trend has materialised in various forms, and one of them is high-stakes testing. Through a Systematic Literature Review (SLR), the present study develops a global examination of this mechanism, in order to understand its scope and influence on the educational system. The results reveal both adverse and beneficial outcomes of highstakes testing and, while the literature highlights unintended consequences, it also stresses the importance of its function as an accountability mechanism in the current society, implying positive aspects. Moreover, findings suggest that viable alternatives for large scale assessments extend beyond high-stakes scenarios.

KEYWORDS LARGE SCALE ASSESSMENTS • HIGH-STAKES TESTING • ALTERNATIVE ASSESSMENTS • SYSTEMATIC LITERATURE REVIEW.

UNA BREVE HISTORIA DE LAS EVALUACIONES DE ALTAS CONSECUENCIAS Y SUS POSIBLES FUTUROS

RESUMEN

El énfasis histórico de la educación en las evaluaciones que promueven el aprendizaje ha evolucionado hacia un modelo competitivo y orientado por el desempeño debido a las reformas educativas y las influencias neoliberales. Esta tendencia se materializó de varias maneras, siendo una las evaluaciones de altas consecuencias (*high-stakes testing*). A través de una Revisión Sistemática de la Literatura (RSL), el presente estudio desarrolla un análisis global de este mecanismo, buscando comprender su alcance e impacto en el sistema educativo. Los resultados revelan resultados tanto adversos como beneficiosos asociados a las evaluaciones de altas consecuencias. Aunque la literatura destaca consecuencias no deseadas, también enfatiza la importancia de su función como mecanismo de responsabilización en la sociedad contemporánea, indicando aspectos positivos. Además, los hallazgos sugieren que las alternativas viables para evaluaciones a gran escala van más allá de los escenarios de altas consecuencias.

PALABRAS CLAVE EVALUACIÓN A GRAN ESCALA • EVALUACIONES DE ALTAS CONSECUENCIAS • EVALUACIÓN ALTERNATIVA • ESTUDIO BIBLIOGRÁFICO.

Recebido em: 11 MARÇO 2024
Aprovado para publicação em: 15 OUTUBRO 2024

INTRODUÇÃO

Ao longo da história da educação, a avaliação tem se mostrado essencial para a aprendizagem. Seu papel principal tem sido, historicamente, um mecanismo para os professores apoiarem a aprendizagem e melhorarem as capacidades e os conhecimentos dos alunos (Hayward, 2015). No entanto, a partir das novas reformas educacionais ocorridas a partir da década de 1990, influenciadas pela Nova Gestão Pública e práticas neoliberais, as avaliações passaram a ser marcadas por mecanismos de competição e desempenho (Verger, Parcerisa et al., 2019), colocando em risco a essência original da avaliação e, consequentemente, da educação. Esse fenômeno, que Ball (2003, 2012c) denomina "performatividade", restringe e fragmenta a aprendizagem (Wyse et al., 2015) e tem implícita uma preocupação excessiva com métricas, medidas e números (Ball, 2015). Assim, obstrui e limita as possibilidades da educação, em vez de expandi-las e enriquecê-las (Ball, 2012a, 2012b). Essas tendências colocam em risco os processos de aprendizagem que deveriam ocorrer nas salas de aula (Madaus & Russell, 2010), além de influenciar a essência dos seres humanos (Ball, 2017).

Atualmente, o que ocorre internacionalmente em muitas salas de aula depende de uma série de fatores que não se baseiam exatamente nas necessidades, interesses e capacidades das crianças que ocupam essas salas (Hoyuelos & Cabanellas, 1996). Em vez disso, são determinados por testes de alto impacto ou avaliações em larga escala (ALEs) em geral e, com eles, processos padronizados (Ball, 2003, 2017; Madaus & Russell, 2010). Assim, as avaliações e o aprendizado que poderiam aprimorar capacidades e criar conhecimento autêntico limitaram os professores a se concentrar na preparação dos alunos para os testes (Berliner, 2011). A razão para isso é que as avaliações se tornaram um fim em si mesmas, com foco na responsabilização (Hayward, 2015) e classificações (Jones & Ennes, 2018). Da mesma forma, os resultados dos testes estão vinculados a julgamentos de desempenho escolar e docente, e, enquanto essa dinâmica permanecer inalterada, a avaliação da aprendizagem e todas as experiências em sala de aula serão determinadas por ela (Hayward, 2015).

Assim, embora os testes de alto impacto desempenhem um papel de responsabilidade que é necessário para a sociedade (Bovens et al., 2008), devido aos dados significativos que fornecem (Schillemans et al., 2013), sua estrutura e os interesses envolvidos estão colocando em risco a educação, as subjetividades das pessoas, a aprendizagem e o significado formativo da educação nos contextos em que ocorrem (Madaus & Russell, 2010; Helfenbein, 2004; Schillemans, 2016; Falabella, 2021). Portanto, o presente estudo explorará os pontos fortes e as limitações desse mecanismo, bem como possíveis formas de melhorar a situação atual. Essa justificativa e preocupação se traduzem em uma questão principal, que será a força motriz desta pesquisa: Como o futuro da avaliação em larga escala pode servir tanto como

um facilitador para o aprimoramento do aprendizado quanto como um mecanismo robusto para garantir a qualidade nos sistemas educacionais e, assim, melhorar as atuais avaliações em larga escala? Isso será realizado por meio de uma Revisão Sistemática da Literatura de Método Misto, que, por um lado, investigará quais têm sido as principais consequências dos testes de alto impacto por meio de uma Revisão de Revisões e, por outro lado, usará uma Revisão de Síntese Interpretativa Crítica para explorar as melhores alternativas ou possíveis futuros para os testes de alto impacto.

HISTÓRIA DOS TESTES DE ALTO IMPACTO

A história dos testes de alto impacto começa com uma mudança significativa na governança ocorrida na década de 1990, inicialmente em algumas nações ocidentais (Murphy, 2021), como os Estados Unidos, o Reino Unido e países europeus (Levi-Faur, 2012; Lynn, 2012). Com o tempo, essa prática se espalhou para outros territórios, como a América Latina (Verger, Fontdevila et al., 2019; Rhodes, 1996). Essa mudança exigiu que os governos alterassem sua forma de governar, passando de uma estrutura centralizada, em termos de poder e controle, para uma descentralizada, que consiste em dar poderes e controle para novas entidades, como mercados, agências ou instituições políticas, governos regionais e organizações não governamentais, entre outras (Levi-Faur, 2012; Rhodes, 2007). Os estudiosos denominaram esse fenômeno de "Nova Governança" (Lynn, 2012).

Essa Nova Governança (Lynn, 2012) foi caracterizada pelo conceito de "governar à distância" (Murphy, 2021, p. 53), marcado por regulamentação, redes de conexão e a criação de padrões (Levi-Faur, 2012). Junto com a expansão do governo e a descentralização, surgiu a "*problématique*" (Levi-Faur, 2012, p. 13). Isso exigiu que, para manter a legitimidade e a eficácia e preservar a qualidade dos serviços (Börzel, 2010; Schillemans et al., 2013; Link & Scott, 2010), essencialmente para demonstrar "boa governança" (Murphy, 2021, p. 33), muitos governos tivessem de optar pelo uso de mecanismos que a academia chamou de "técnicas gerenciais" (Ackerman, 2004; Hood & Dixon, 2016). Esses mecanismos ou estratégias ficaram conhecidos como "nova gestão pública, ou NGP, para abreviar" (Murphy, 2021, p. 42).

Uma das medidas para salvaguardar a "boa governança" é o teste de alto impacto (Nichols & Berliner, 2007), que é um "instrumento de política" (Levi-Faur, 2012) destinado a medir o "aprendizado" ou o desempenho de alunos, professores e escolas por meio de testes padronizados para avaliar a educação e assegurar sua qualidade (Nichols & Berliner, 2007; Muller, 2018). Além disso, esse tipo de avaliação é caracterizado pelo fato de possuir "alto impacto" (Jones & Ennes, 2018), pois exerce impactos cruciais sobre os agentes educacionais, como a promoção de

educadores e alunos, o ajuste de salários e a alocação de recursos (Jones & Ennes, 2018; Gregory & Clarke, 2003).

Naturalmente, há várias causas e fatores que influenciam a consolidação dos testes de alto impacto. Juntamente com a evolução da governança (Levi-Faur, 2012), há outro aspecto envolvido, a introdução da educação de massa da Europa Ocidental no século XIX. Isso envolveu a incorporação e a expansão da educação geral obrigatória e controlada pelo Estado (Soysal & Strang, 1989), que, segundo alguns autores, estava ligada ao desenvolvimento e crescimento econômico (Zinkina et al., 2016; Westberg et al., 2019). Esse fenômeno histórico incluiu os Estados Unidos (Beadie, 2019) e depois, no final do século XIX e início do século XX, foi introduzido na América Latina (Frankema, 2009), Austrália, Nova Zelândia e, em menor escala, na Ásia e na África (Zinkina et al., 2016).

A educação em massa foi crucial para a formação dos sistemas educacionais nacionais e, portanto, para a construção e unificação de suas políticas nacionais (Ramirez & Boli, 1987); ela é considerada "um subproduto da industrialização" (Green, 2013, p. 47). Entre os séculos XIX e XX, ocorreram muitas mudanças nos sistemas educacionais devido à educação em massa, algumas das quais foram a unificação do currículo, a regulamentação dos requisitos de entrada em diferentes níveis do sistema e, o que é mais relevante para o presente estudo, a educação se concentrou na reprodução, mecanização e memorização (Benavot et al., 1991). Posteriormente, surgiram as medidas de avaliação nacional (Green, 2013) e a sistematização foi essencial para demonstrar habilidades e distinguir as pessoas, dando destaque à meritocracia e à competência (Green, 2013).

Além disso, os sistemas de educação em massa exigiam que os Estados assumissem um papel fundamental na educação: não apenas no financiamento, mas também na regulamentação e administração, o que, como mencionado anteriormente, deu origem a estruturas nacionais de currículo e avaliação (Ramirez & Boli, 1987). Desde então, esses sistemas de avaliação evoluíram e, em meados do século XIX, as escolas começaram a aplicar testes padronizados baseados puramente na memorização (predominantemente orais) (Huddleston & Rockwell, 2015). Com a convergência de influências internacionais e eventos históricos, surgiram as avaliações nacionais. Posteriormente, com a Segunda Guerra Mundial, sistemas internacionais de avaliação foram formalizados e disseminados (Kamens & McNeely, 2010). Embora o momento de adoção dessas dinâmicas varie entre os países, e alguns ainda não tenham adotado esses mecanismos, as avaliações em larga escala (ALEs) e os testes de alto impacto (TAIs) têm ampliado cada vez mais seu alcance, nacional e internacionalmente, influenciando e moldando-se mutuamente (Verger, Parcerisa et al., 2019).

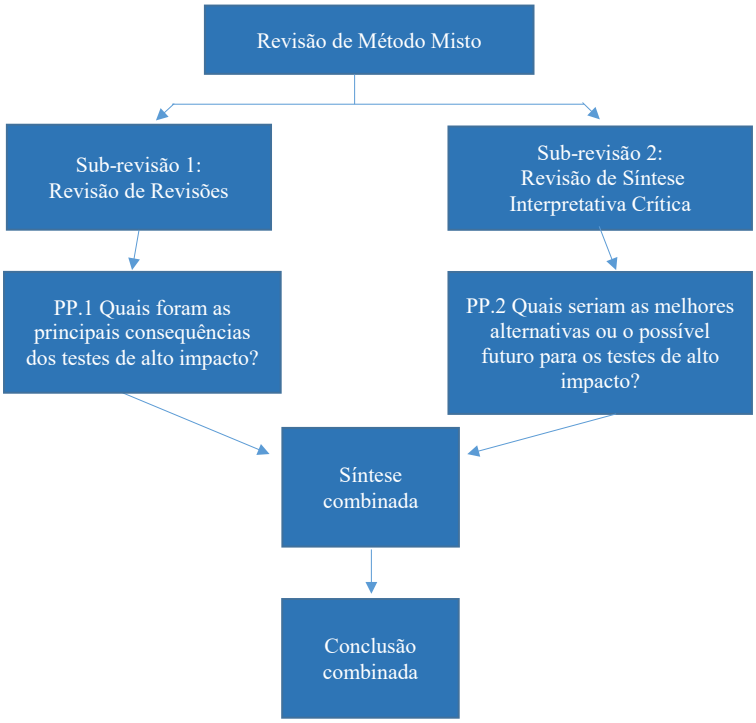
Mais adiante, essas avaliações em larga escala e instrumentos de políticas começaram, progressivamente, a assumir estruturas mais padronizadas como resultado de reformas e políticas educacionais (Ball, 2003). Com o tempo, foram assumindo funções que causam impactos significativos sobre indivíduos e instituições, fazendo com que esses exames passassem a ter consequências mais relevantes (Nichols & Berliner, 2007). Assim, o desempenho passou a determinar a concessão de diplomas, o acesso aos próximos níveis de educação, os salários dos professores e, o que é mais importante, essa nova tendência criou classificações e tabelas classificatórias nas quais se baseiam muitas reputações e bilhões de libras (Zhao, 2014). Como resultado, números, estatísticas, padrões, medições e performatividade tornaram-se formas dominantes na sociedade moderna e, com eles, os testes de alto impacto (Ball, 2003; Gregory & Clarke, 2003; Zhao, 2014; Muller, 2018).

METODOLOGIA

A metodologia utilizada no presente estudo é a Revisão Sistemática da Literatura (RSL), que fornece uma visão geral sobre o que é conhecido acerca de um tema específico (Gough et al., 2013). Particularmente, trata-se de uma RSL de Método Misto, pois "possui sub-revisões que formulam perguntas sobre diferentes aspectos de uma questão" (Gough et al., 2012, p. 6, tradução nossa). A primeira sub-revisão foi uma Revisão de Revisões (Gough et al., 2017) e teve como objetivo responder quais foram as principais consequências dos TAIs. Para fins de clareza, essa pergunta de pesquisa foi designada como número 1 (PP.1). A segunda sub-revisão foi uma Revisão de Síntese Interpretativa Crítica, que resume dados, desafia e questiona pressupostos geralmente aceitos (Dixon-Woods et al., 2006, p. 4), buscando responder quais seriam as melhores alternativas ou o possível futuro para os testes de alto impacto (TAIs). Essa pergunta de pesquisa foi designada como número 2 (PP.2). Ambas as revisões têm como objetivo responder a uma questão principal, a saber: como o futuro das avaliações em larga escala pode servir tanto como um facilitador para a melhoria da aprendizagem quanto como um mecanismo robusto para garantir a qualidade nos sistemas educacionais e, assim, aprimorar as atuais ALEs?

Esta RSL utilizou uma metodologia mista, uma vez que os procedimentos para encontrar padrões nos artigos, analisar os dados e criar códigos foram qualitativos; e os procedimentos para associar códigos aos padrões e depois quantificá-los em um gráfico foram quantitativos. Para uma compreensão mais clara da metodologia, a Figura 1, baseada em Gough et al. (2012), apresenta uma representação da estrutura.

FIGURA 1
Estrutura da Revisão de Método Misto realizada neste estudo, com base na proposta de Gough et al. (2012)



Fonte: Elaboração da autora (2024).

Etapas da Revisão Sistemática da Literatura

As etapas que foram seguidas para a realização dessa RSL de Método Misto baseiam-se na proposta da University College London (UCL) (2023) e de Gough et al. (2013). Na primeira sub-revisão, foram selecionados 15 artigos; na segunda sub-revisão, foram selecionados 25 artigos. Os comandos usados para as buscas estão detalhados nos apêndices A e B. A segunda sub-revisão exigiu 25 buscas, vários sinônimos e a leitura de 200 resumos. No entanto, quando essa segunda busca resultou em um número escasso de artigos relevantes para a questão, foi usada a estratégia de Gough et al. (2013) de consultar especialistas. Assim, o professor Clive Dimmock e a dra. Clara Fontdevila forneceram o que, de acordo com seus critérios, eram artigos relevantes para a pesquisa. Mais detalhes podem ser consultados no Apêndice C.

Os critérios de inclusão e exclusão usados para a seleção dos estudos estão detalhados abaixo, na Tabela 1. É importante observar que a revisão da literatura considerou exclusivamente publicações escritas em inglês.

TABELA 1
Critérios baseados em Gough et al. (2012) e UCL (2023)

SUB-REVSÃO	CRITÉRIOS DE INCLUSÃO	CRITÉRIOS DE EXCLUSÃO	CRITÉRIOS DE INCLUSÃO (EM COMUM)	CRITÉRIOS DE EXCLUSÃO (EM COMUM)
1) Revisão de Revisões	• Apenas Revisões Sistemáticas da Literatura foram selecionadas	• Todos os artigos não relacionados aos efeitos ou consequências de testes de alto impacto (ou outros sinônimos)	• Alcance geográfico: internacional • Idioma: inglês • Os artigos foram selecionados com base em sua relevância (capacidade de resposta) para as perguntas de pesquisa correspondentes • Escopo de tempo: de 1990 até os dias atuais, 2023 • Devem conter considerações éticas válidas (Petticrew & Roberts, 2006) • Apenas artigos revisados por pares • Apenas estudos focados em testes de alto impacto para os alunos • Os artigos poderiam considerar os efeitos sobre crianças e professores	• Trabalhos focados em exames de conclusão de curso • Focados em educação terciária ou superior • Focados nos impactos dos testes computadorizados e na avaliação de professores
2) Revisão de Síntese Interpretativa Crítica	• Todos os tipos de coleta de dados (qualitativas e quantitativas) • Os estudos devem ter uma metodologia robusta e consideração ética detalhada			

Fonte: Elaboração da autora (2024).

Avaliação da qualidade

Em primeiro lugar, o principal critério para garantir a qualidade foi a relevância dos estudos selecionados em relação às perguntas de pesquisa (Gough et al., 2013). Em segundo, considerou-se o rigor metodológico (Gough et al., 2013). Em terceiro, foram selecionados apenas artigos revisados por pares. Por fim, em quarto, apenas artigos de periódicos foram incluídos, excluindo-se outros tipos de documentos, como *blogs* ou literatura cinzenta, a fim de assegurar a confiabilidade, uma vez que tais documentos podem apresentar vieses, ser incompletos e metodologicamente desafiadores de avaliar (Hopewell et al., 2005).

Síntese e análise de dados

O presente estudo examinou os dados, pesquisou padrões neles, criou códigos e atribuiu valores a esses códigos. Posteriormente, eles foram organizados em gráficos e interpretados (Gough et al., 2013). Para converter os dados de qualitativos em quantitativos, as duas análises foram diferentes. Na sub-revisão 1, como ela

mede as consequências e o seu peso, foram associadas pontuações diferentes com base no rigor e na robustez dos artigos. Para medir isso, foram selecionados os cinco critérios de qualidade a seguir:

1) Número de estudos revisados.

2) Apresenta critérios que asseguram a qualidade.

3) Descreve o banco de dados utilizado.

4) Apresenta as palavras usadas para a pesquisa.

5) Indica os critérios de inclusão e exclusão.

Assim, de acordo com o número de critérios que cada artigo preenchia, uma pontuação era atribuída a ele. Essas pontuações, quando associadas aos códigos de consequência, forneceriam uma pontuação final e somativa e, portanto, um peso para cada artigo. Os detalhes da atribuição de pontuação são fornecidos a seguir:

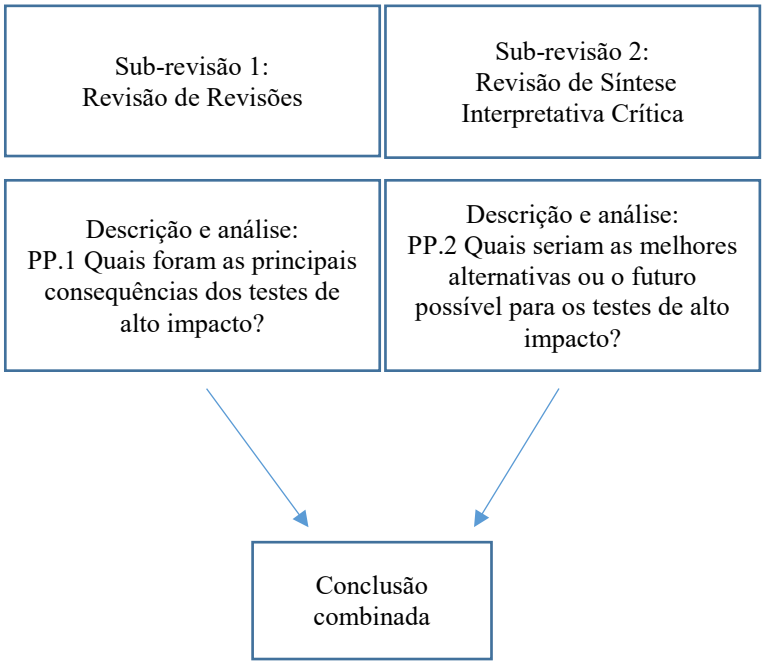
- 3 pontos: se atender a 4 ou mais critérios.
- 2 pontos: se atender a 3 critérios.
- 1 ponto: se atender a 2 ou menos dos critérios acima.

Da mesma forma, na sub-revisão 2, como o que era essencial eram as próprias propostas, e não a robustez da pesquisa, o que importa é a ideia em si. Portanto, todas elas possuem a mesma validade.

RESULTADOS

Os resultados serão explicados para cada pergunta de pesquisa, separadamente. Por fim, ambas as perguntas de pesquisa serão combinadas na discussão e na conclusão. Essa organização está ilustrada na Figura 2.

FIGURA 2
Organização da apresentação e análise dos resultados



Fonte: Elaboração da autora (2024).

Resultados PP.1 – Quais foram as principais consequências dos testes de alto impacto?

Para identificar as consequências predominantes na literatura existente e determinar os artigos que forneceram suporte substancial e evidência empírica, todos os artigos foram avaliados de acordo com sua robustez. O gráfico apresentado na Figura 3, a seguir, mostra a frequência de cada consequência mencionada nos artigos e, além disso, atribui uma pontuação de 1 a 3 que reflete o nível de rigor e robustez. Essas pontuações são baseadas nos critérios descritos na seção “Síntese e análise de dados”.

FIGURA 3

Gráfico obtido da sub-revisão 1: Revisão das Revisões focada em responder à PP.1 – Quais foram as principais consequências dos testes de alto impacto?

Fonte: Elaboração da autora (2024).

Os resultados da Revisão das Revisões sugerem que, entre as consequências com as pontuações mais altas e as repetidas com mais frequência, estão: estreitamento do currículo (Acosta et al., 2020; Au, 2009; Sigvardsson, 2017; Boon et al., 2007; Verger, Fontdevila et al., 2019; Verger, Parcerisa et al., 2019; Anderson, 2012; Cimbricz, 2002; Emler et al., 2019; Harlen & Crick, 2003), colocando em risco a equidade (Acosta et al., 2020; Au, 2009; Boon et al., 2007; Bacon & Pomponio,

2023; Hamilton et al., 2013; Verger, Parcerisa et al., 2019; Anderson, 2012; Emler et al., 2019; Lee, 2008), limitação ou restrição de habilidades de ordem superior (Acosta et al., 2020; Au, 2009; Verger, Fontdevila et al., 2019; Verger, Parcerisa et al., 2019; Anderson, 2012, Emler et al., 2019), professores que ensinam para o teste (Acosta et al., 2020; Au, 2009; Hamilton et al., 2013; Cimbricz, 2002; Ehren et al., 2016; Emler et al., 2019; Nichols, 2007; Harlen & Crick, 2003), aprendizado prejudicado pelos TAIs (Boon et al., 2007; Hamilton et al., 2013; Anderson, 2012; Cimbricz, 2002; Emler et al., 2019; Nichols, 2007) e a confiabilidade questionada desses testes (Acosta et al., 2020; Verger, Fontdevila et al., 2019; Hamilton et al., 2013; Cimbricz, 2002; Lee, 2008; Nichols, 2007).

Mais detalhadamente, as consequências negativas após as codificações resultaram em um total de 30. Códigos semelhantes foram organizados em 6 temas: (1) o currículo e o que acontece na sala de aula; (2) como eles influenciaram os professores; (3) a consistência e a confiabilidade dos testes de alto impacto; (4) a cultura escolar; (5) a equidade; e (6) a influência na subjetividade das pessoas. Em primeiro lugar, os estudos revisados sugerem que a educação foi afetada, porque os testes de alto impacto levaram a um estreitamento curricular (Acosta et al., 2020; Au, 2009; Boon et al., 2007; Verger, Fontdevila et al., 2019; Verger, Parcerisa et al., 2019; Anderson, 2012; Cimbricz, 2002; Emler et al., 2019; Harlen et al., 2002). Além disso, essa redução pressupõe uma visão do currículo em seu amplo espectro, considerando o ensino, o conteúdo e as habilidades que são desenvolvidas, bem como as interações que ocorrem (e não ocorrem) dentro da sala de aula. Um exemplo concreto disso é apresentado por Au (2009), com relação à situação nos Estados Unidos: "71% dos distritos relataram ter cortado pelo menos uma matéria para aumentar o tempo gasto com leitura e matemática como resposta direta aos testes de alto impacto exigidos pela NCLB" (Renter et al., 2006, como citado em Au, 2009, p. 46, tradução nossa).

Além disso, vários autores destacam que as habilidades de ordem superior, como o pensamento crítico, divergente e criativo, não têm prioridade na sala de aula (Acosta et al., 2020; Au, 2009; Verger, Fontdevila et al., 2019; Verger, Parcerisa et al., 2019; Anderson, 2012; Cimbricz, 2002; Emler et al., 2019); assim, o ensino é reduzido (Bacon & Pomponio, 2023; Ehren et al., 2016; Emler et al., 2019) à mera repetição e memorização de fatos (Au, 2009). Cada vez mais, os professores têm se concentrado em preparar os alunos para os testes (Emler et al., 2019; Nichols, 2007) e, consequentemente, as matérias que não são avaliadas nos testes de alto impacto (Harlen & Crick, 2003), como arte, música, esportes, poesia, entre outras, têm tido sua importância diminuída (Au, 2007), culminando, portanto, em prejuízo do aprendizado (Boon et al., 2007; Hamilton et al., 2013; Anderson, 2012; Cimbricz, 2002, Emler et al., 2019; Nichols, 2007). Acosta et al. (2020, p. 536, tradução nossa) destacam a

opinião de um aluno: "Estamos aprendendo apenas o conteúdo dos testes e não o que deveríamos saber e ir para a faculdade".

Em segundo lugar, essas consequências negativas também afetaram os professores, sua profissão e seu profissionalismo. Com base em fortes recompensas e sanções, os docentes passaram a ter atitudes relacionadas à manipulação de resultados ou fraude (Ehren et al., 2016; Emler et al., 2019; Hamilton et al., 2013). O TAI levou vários educadores a deixar suas escolas (Boon et al., 2007), aumentou sua sobrecarga de trabalho (Ehren et al., 2016) e competitividade (Verger, Parcerisa et al., 2019) e colocou em risco o bem-estar dos professores (Anderson, 2012; Cimbricz, 2002; Ehren et al., 2016; Emler et al., 2019; Harlen & Crick, 2003). Ele faz com que os professores contrariem suas crenças e valores (Ehren et al., 2016), percam a motivação (Hamilton et al., 2013; Emler et al., 2019) e corrói a essência das disciplinas e do ensino (Sigvardsson, 2017; Anderson, 2012), entre outros elementos apresentados na Figura 3. Além disso, junto com os danos aos professores, há o fenômeno da desprofissionalização desses trabalhadores, que implica a perda da autonomia profissional (Verger, Fontdevila et al., 2019; Anderson, 2012; Cimbricz, 2002; Ehren et al., 2016; Emler et al., 2019). Isso também contribuiu para comprometer a qualidade dos professores (Verger, Parcerisa et al., 2019; Anderson, 2012).

Em terceiro lugar, o TAI falha na consistência das informações que fornece sobre o aprendizado; 6 dos 15 estudos (com uma pontuação total de 12) declaram que a confiabilidade dos mecanismos é questionável (Cimbricz, 2002; Lee, 2008; Nichols, 2007; Hamilton et al., 2013; Verger, Fontdevila et al., 2019; Acosta et al., 2020); 4 artigos com um total de 7 pontos mencionam que os estudos sobre benefícios são inconclusivos (Hamilton et al., 2013; Verger, Parcerisa et al., 2019; Lee, 2008; Nichols, 2007); e 1 estudo com 3 pontos, portanto um estudo forte, declara que a medição de desempenho e as pontuações refletem resultados que não estão diretamente relacionados ao aprendizado e ao conhecimento (Boon et al., 2007). As informações são restritas, pois são limitadas, na maioria dos casos, por perguntas de múltipla escolha ou perguntas fechadas. Nesse cenário, o teste de alto impacto não é capaz de fornecer informações suficientes para avaliar o aprendizado e sua complexidade (Boon et al., 2007; Acosta et al., 2020).

Em quarto lugar, estudos sugerem que o TAI gerou uma cultura competitiva nas escolas e no sistema educacional em geral (Verger, Fontdevila et al., 2019), em que os professores até se voltaram contra alguns alunos com desempenho ruim, marginalizando-os (Bacon & Pomponio, 2023; Hamilton et al., 2013). Como resultado, a educação, os docentes, o aprendizado e a própria pedagogia foram afetados (Au, 2009; Hamilton et al., 2013; Anderson, 2012; Cimbricz, 2002; Emler et al., 2019).

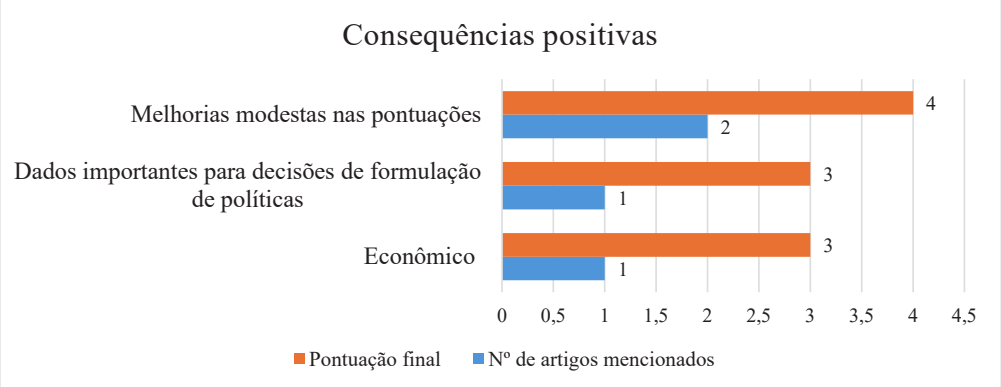
Em quinto lugar, temos a equidade, que foi mencionada em 9 dos 15 artigos (o que é substancial em relação às outras consequências). Esses estudos indicam

que a equidade está em risco devido ao TAI, porque os testes tendem a perpetuar a injustiça e as desigualdades (Emler et al., 2019; Bacon & Pomponio, 2023; Harlen & Crick, 2003) que afetam os alunos com deficiências, os que estão aprendendo inglês e os de comunidades desfavorecidas (Boon et al., 2007). Estudos também sugerem que o TAI aumenta as disparidades raciais e socioeconômicas (Acosta et al., 2020; Emler et al., 2019). Emler et al. (2019, p. 589, tradução nossa) fornecem mais detalhes, observando que

> . . . as ALEs têm consistentemente revelado grandes lacunas nos resultados entre diferentes grupos de alunos. Essas lacunas são, em grande parte, resultado de desigualdades socioeconômicas e raciais, além de outros fatores que estão fora do controle de escolas e professores. . . . Em outras palavras, os esforços para reduzir a lacuna de desempenho têm ampliado a lacuna de oportunidades, criando mais desigualdade e injustiça.

Sexto, uma das consequências mais marcantes dos testes de alto impacto é a erosão da complexidade humana dos alunos, professores e da própria educação (Emler et al., 2019; Acosta et al., 2020; Bacon & Pomponio, 2023; Ehren et al., 2016). Devido à padronização e homogeneização (Emler et al., 2019), ao estreitamento do currículo e do ensino (Acosta et al., 2020; Boon et al., 2007), à prevalência de métodos de aprendizagem passivos (Anderson, 2012) e à restrição a habilidades cognitivas de baixo nível, como memorização e repetição (Au, 2009), a subjetividade, a diversidade e a individualização têm sido colocadas em risco. Isso é mencionado em quatro estudos, com uma pontuação total de sete (Au, 2009; Emler et al., 2019; Harlen & Crick, 2003; Hamilton et al., 2013). Ainda, outros estudos declaram que os testes de alto impacto limitam a possibilidade de que todos os alunos tenham sucesso (Acosta et al., 2020) e restringem o desenvolvimento de talentos e paixões específicos (Emler et al., 2019; Harlen & Crick, 2003).

FIGURA 4
Gráfico obtido da sub-revisão 1: Revisão das Revisões voltadas para responder à PP.1 – Quais foram as principais consequências dos testes de alto impacto?



Fonte: Elaboração da autora (2024).

As 3 consequências positivas identificadas na Figura 4 foram encontradas por meio da análise desses 15 artigos. Essas consequências foram extraídas de 4 textos diferentes, 3 dos quais são textos de 3 pontos e, portanto, são estudos fortes, rigorosos e robustos. Isso demonstra que essas implicações são legítimas, importantes e devem ser consideradas (Verger, Fontdevila et al., 2019; Verger, Parcerisa et al., 2019; Lee, 2008; Anderson, 2012).

Conforme mostrado na Figura 4, esses estudos apresentam consequências positivas do teste de alto impacto relacionadas a elementos estruturais e funcionais. Eles se concentram sobretudo nos elementos políticos, na elaboração de políticas e nas implicações financeiras (Verger, Fontdevila et al., 2019; Verger, Parcerisa et al., 2019), que visam à transparência e à boa governança, e indicam que os testes de alto impacto respondem a uma necessidade, especialmente em termos de responsabilidade (Anderson, 2012). Além disso, é importante observar que os atuais mecanismos de avaliação em larga escala são baratos em comparação com outras opções e, portanto, atraentes para a formulação de políticas em diferentes países (Verger, Fontdevila et al., 2019).

Outra consequência é um aumento moderado nas pontuações, presente em 2 artigos (Lee, 2008; Anderson, 2012). Essa questão é controversa, pois levanta a dúvida sobre como essas pontuações cresceram ou o que elas realmente significam. Esses aumentos nas pontuações podem estar relacionados a um dos fenômenos apresentados nas consequências negativas, como fraudes, ensino voltado para o teste, estreitamento do currículo, entre outros (Hamilton et al., 2013; Boon et al., 2007). Há elementos críticos que enfraquecem o argumento a favor da melhoria no desempenho, como, por exemplo, as questões sobre sua confiabilidade (6 artigos, pontuação de 12) (Verger, Fontdevila et al., 2019) e os efeitos benéficos inconclusivos apontados por vários estudos (4 com pontuação de 7) (Verger, Parcerisa et al., 2019; Hamilton et al., 2013). A realidade é que há resultados contraditórios.

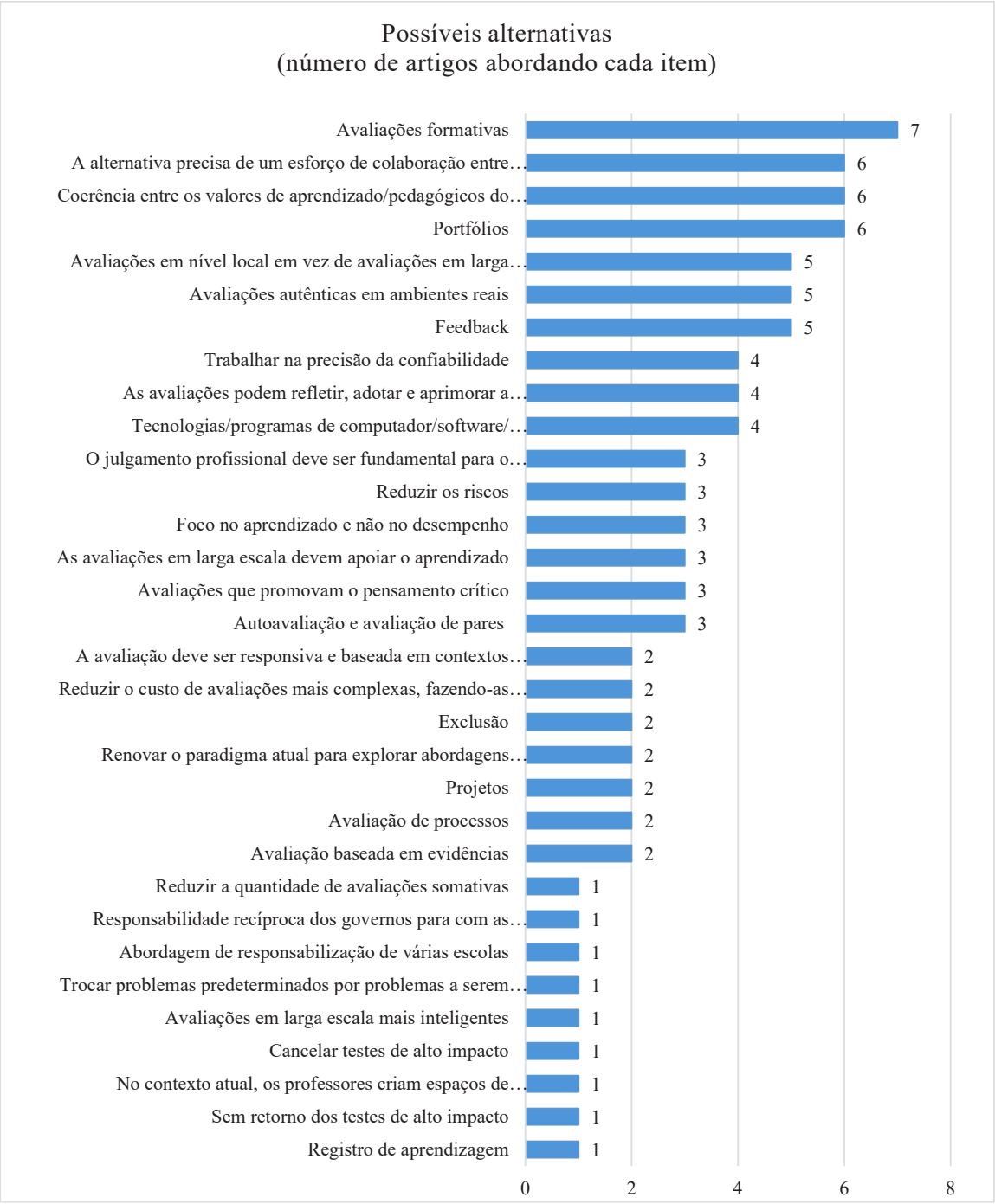
Em resumo, o TAI tem pontos fortes que devem ser considerados quando se pensa em possibilidades futuras para esses tipos de avaliação. Embora o TAI seja capaz de responder a uma necessidade válida de objetividade, esses mecanismos têm consequências significativas e não intencionais com efeitos negativos sobre a educação, os professores, os alunos, seu aprendizado e suas individualidades, além de comprometer fatores éticos e de equidade. Portanto, para o futuro, é necessário conciliar os pontos fortes com a melhoria dos pontos fracos.

Resultados PP.2 – Quais seriam as melhores alternativas ou o possível futuro para os testes de alto impacto?

A compilação a seguir representa uma amostra do extenso corpo de literatura sobre o assunto. Os resultados apresentados são baseados na revisão dos 25 artigos selecionados.

FIGURA 5

Gráfico obtido da sub-revisão 2: Revisão da Síntese Interpretativa Crítica com foco na resposta à PP.2 – Quais seriam as melhores alternativas ou o possível futuro para os testes de alto impacto?



Fonte: Elaboração da autora (2024).

A maioria dos resultados apresentados na Figura 5 oferece ideias e possibilidades de melhorias para os métodos atuais de avaliação em larga escala. No entanto, dois artigos sugerem abandoná-los completamente, optando pela exclusão dos alunos e das escolas (Wang, 2017; Ashadi et al., 2022) ou pelo cancelamento das avaliações (Ashadi et al., 2022). Essas opções, embora extremas, apresentam uma

posição que, em relação ao que é mencionado anteriormente, atendem a uma necessidade atual e possuem pontos fortes relevantes ao se considerar possíveis futuros.

Por outro lado, o restante da literatura revisada reconhece o valor desses mecanismos para coletar e fornecer informações sobre a aprendizagem, a qualidade da educação, atender à necessidade de responsabilização das instituições educacionais públicas e possibilitar a obtenção de dados sobre processos que promovem discussões aprofundadas sobre o que ocorre em sala de aula (Dorn, 1998; Chudowsky & Pellegrino, 2003). Da mesma forma, os TAIs funcionam como uma ferramenta de seleção e alocação de alunos em momentos educacionais cruciais, como a transição do ensino primário para o secundário, o avanço do ensino escolar para o superior e a tomada de decisões sobre a distribuição de recursos (Suto & Oates, 2021). Além disso, eles indicam a necessidade urgente de melhorar, repensar os atuais TAIs e reconsiderar o paradigma vigente (Chudowsky & Pellegrino, 2003; Volante, 2007). Caso contrário, as consequências negativas e suas fragilidades continuarão a impactar negativamente a aprendizagem e perpetuar desigualdades (Volante, 2007; Lingard, 2009; Syverson, 2011).

Os resultados da presente análise foram organizados em três grupos. Primeiro, alternativas concretas que podem oferecer soluções mais simples; segundo, mudanças sistêmicas que envolvem alternativas mais complexas e profundas; e, terceiro, sugestões sobre como os processos de desenvolvimento de novas alternativas para os testes de alto impacto podem ser conduzidos.

Em primeiro lugar, os autores revisados sugerem: a incorporação do *feedback* (Cato & Walker, 2022; Chudowsky & Pellegrino, 2003; Brown et al., 2014; Zimmerman & Dibenedetto, 2008; Beyond Test Scores Project [BTS Project] & National Education Policy Center [NEPC], 2023); a consideração e a combinação da avaliação em contextos autênticos e espontâneos (Syverson, 2011; Roberson, 2011; Brown et al., 2014; Volante, 2007; Lingard, 2009); a avaliação de processos (Behizadeh & Lynch, 2017; Zimmerman & Dibenedetto, 2008); a autoavaliação e avaliação de pares (Açıkalın, 2014; Chudowsky & Pellegrino, 2003; Roberson, 2011); projetos (Açıkalın, 2014; BTS Project & NEPC, 2023); maior ênfase em avaliações formativas para reduzir a proeminência das avaliações somativas (Açıkalın, 2014; Chudowsky & Pellegrino, 2003; Roberson, 2011; Brown et al., 2014; Gillanders et al., 2021; Zimmerman & Dibenedetto, 2008; Hutchinson & Hayward, 2005; Hayward et al., 2004; Hayward & Spencer, 2010; BTS Project & NEPC, 2023); incentivo ao pensamento crítico e habilidades de ordem superior (Ab Kadir, 2017; Roberson, 2011; Brown et al., 2014); foco no aprendizado e não no desempenho (Volante, 2007; Lingard, 2009; BTS Project & NEPC, 2023); e a criação de espaços para formas de avaliação não predeterminadas (Beghetto, 2019).

Mais concretamente, alguns autores recomendam: portfólios (Syverson, 2011; Açıkalın, 2014; Chudowsky & Pellegrino, 2003; Behizadeh & Lynch, 2017; Herman

& Winters, 1994; BTS Project & NEPC, 2023) e avaliações em larga escala mais inteligentes por meio do uso de tecnologias, programas de computador, *softwares* e inteligência artificial que permitam a coleta de informações de forma formativa, proporcionando *feedback* e incorporando muitos dos elementos mencionados anteriormente (Beghetto, 2019; Chudowsky & Pellegrino, 2003; Behizadeh & Lynch, 2017).

Em segundo lugar, em relação às alternativas sistêmicas e complexas, os estudos propõem que as políticas devem buscar coerência entre currículo e avaliação, e que ambos devem estar alinhados com os propósitos pedagógicos e de aprendizagem (Ab Kadir, 2017; Dorn, 1998; Chudowsky & Pellegrino, 2003; Volante, 2007; Zimmerman & Dibenedetto, 2008; BTS Project & NEPC, 2023). Seguindo a mesma lógica, vários artigos apontam que é necessário que esses mecanismos promovam e abordem a complexidade e a abrangência dos seres humanos e da educação (Roberson, 2011; Volante, 2007; Hayward & Spencer, 2010; BTS Project & NEPC, 2023). Além disso, a literatura sugeriu que as avaliações em larga escala deveriam estar mais focadas em níveis locais (Dorn, 1998; Gillanders et al., 2021; Moss, 2022; Ashadi et al., 2022; Volante, 2007; BTS Project & NEPC, 2023).

Vários estudos enfatizaram a importância de devolver a responsabilidade e a confiança no papel profissional dos professores, restaurando sua autonomia profissional (Hutchinson & Hayward, 2005; Hayward & Spencer, 2010; Lingard, 2009). Ademais, os autores mencionaram que o maior problema com os TAIs está em seus riscos, e, portanto, independentemente da decisão tomada com base em melhorias ou alternativas, a forma de avaliação deve reduzir esses riscos (Behizadeh & Lynch, 2017; Hooge et al., 2012; BTS Project & NEPC, 2023).

Em terceiro lugar, em relação a como realizar o processo de transformação, existem dois elementos cruciais. O primeiro é a necessidade de uma abordagem colaborativa, envolvendo todas as partes interessadas e especialistas (Chudowsky & Pellegrino, 2003; Volante, 2007; Behizadeh & Lynch, 2017; Hutchinson & Hayward, 2005; Hooge et al., 2012; BTS Project & NEPC, 2023). O segundo é o fato de repensar e adotar novos paradigmas, diferentes do pensamento atual, indo além dos limites do sistema (Chudowsky & Pellegrino, 2003; Volante, 2007).

Por fim, Syverson (2011, p. 4, tradução nossa) ilustra perfeitamente o direcionamento necessário para o possível futuro dos testes de alto impacto, afirmando que "os testes padronizados têm se concentrado em padronizar o conteúdo do que é avaliado, em vez de padronizar a estrutura na qual diversos tipos de evidências de aprendizagem podem ser coletados, organizados, compreendidos e avaliados".

DISCUSSÃO

O presente estudo constatou que os testes de alto impacto têm, de fato, consequências negativas e substancialmente prejudiciais (Verger, Fontdevila et al., 2019). Contudo,

ao mesmo tempo, oferecem uma oportunidade e constituem uma forma necessária de transparência e democracia, fundamentais no mundo atual (Schillemans et al., 2013; Bovens, 2010). Além disso, conforme destaca Hayward (2015, p. 38), a avaliação deve ser "como, para e de" aprendizagem; assim, os atores envolvidos na educação têm a oportunidade de reverter e melhorar a situação, transformando esses mecanismos em benefício da aprendizagem e da educação. No entanto, isso exige a consideração de dois elementos: a colaboração, que permite que as pessoas enxerguem o que não pode ser visto individualmente e encontrem soluções ainda não consideradas (Chudowsky & Pellegrino, 2003; Volante, 2007; Hutchinson & Hayward, 2005; Hooge et al., 2012; BTS Project & NEPC, 2023), e a disposição dos formuladores de políticas para analisar novas estruturas e paradigmas (Volante, 2007; Chudowsky & Pellegrino, 2003).

Ampliando a ideia anterior, exemplos concretos aparecem na segunda sub-revisão, demonstrando que novos movimentos e possibilidades têm surgido. Nos Estados Unidos, algumas pessoas e escolas optaram por sair dos sistemas de TAI; elas criaram o movimento "FairTest" (Syverson, 2011). Da mesma forma, na Escócia, iniciou-se um projeto de avaliação formativa, acompanhado pelo governo em parceria com a academia e escolas (Hutchinson & Hayward, 2005; Hayward et al., 2004; Hayward & Spencer, 2010). Seguindo a mesma lógica, as soluções e opções apresentadas nos estudos geram alternativas simples, concretas e significativas, como, por exemplo, a redução do alto impacto atribuída aos testes (Hooge et al., 2012; BTS Project & NEPC, 2023). Este é um pequeno passo, mas que poderia ter efeitos significativos nas salas de aula, nas escolas e na aprendizagem, possivelmente reduzindo algumas das consequências indesejadas.

Por outro lado, abordar apenas aspectos superficiais, em vez de enfrentar a raiz do problema, apenas manterá a questão existente. Por exemplo, concentrar-se exclusivamente em conteúdos ou habilidades específicas como solução. No caso do Pisa, que incorporou a criatividade em suas avaliações, isso poderia levar países e escolas a adotarem comportamentos antiéticos, como ensinar para o teste e restringir currículos para priorizar essa nova habilidade (Beghetto, 2019). Portanto, alcançar melhorias significativas e profundas exige que os problemas sejam tratados em sua origem (Beghetto, 2019).

De forma semelhante, outra solução bastante realista é implementar diferentes formas de avaliação (Hooge et al., 2012), complementando a avaliação somativa com a formativa e facilitando-a por meio de avaliações em grupo por território (Volante, 2007; BTS Project & NEPC, 2023). Dessa forma, o custo não seria excessivo, e o alto esforço necessário poderia ser dividido em pequenas etapas. Além disso, a incorporação de tecnologias no dia a dia das salas de aula permite a avaliação de processos e o fornecimento de *feedback* sob uma perspectiva mais lúdica e

pedagógica, o que também é uma alternativa razoável e eficaz (Behizadeh & Lynch, 2017; Beghetto, 2019). Finalmente, há a proposta de pensar em formas de avaliação sem ideias predeterminadas, com possibilidades abertas de questões ou exercícios, permitindo que os alunos formulem suas próprias ideias, sem que os avaliadores tenham uma ideia predefinida de uma resposta correta (Beghetto, 2019).

Por fim, é relevante considerar que a educação não apenas molda o próprio sistema, mas também influencia a sociedade e, consequentemente, a humanidade: "o que os alunos devem saber e o que eles não sabem são fatores altamente controlados pelos exames" (Emler et al., 2019, p. 281, tradução nossa). As consequências dessa situação tornam-se mais evidentes quando os indivíduos não têm consciência de suas lacunas de conhecimento. Assim, eles não são capazes nem mesmo de perceber que existem coisas que não sabem, mas este é um tema para investigações futuras. No entanto, enquanto as avaliações de alto impacto persistirem como estão, o conteúdo ensinado nas salas de aula será limitado ao que é avaliado. Tudo que estiver fora desse espectro permanecerá desconhecido ou inexplorado (Emler et al., 2019).

CONCLUSÃO

De forma geral, o presente estudo abordou inicialmente a história dos TAIs, conceitos relacionados e fenômenos associados, com o objetivo de facilitar a compreensão e permitir que o leitor contextualize o tema. Em seguida, a metodologia aprofundou-se nas duas sub-revisões que constituíram a RSL, possibilitando uma investigação mais profunda. A etapa subsequente envolveu a apresentação dos resultados obtidos, seguida da discussão, implicações e recomendações.

Portanto, é necessário retornar à principal questão de pesquisa, apresentada no início deste estudo: como o futuro da avaliação em larga escala pode servir tanto como um facilitador para a melhoria da aprendizagem quanto como um mecanismo robusto para garantir a qualidade nos sistemas educacionais? A resposta é composta por algumas ideias principais. Primeiramente, há a necessidade de parcerias colaborativas que levem em consideração tanto os alunos quanto os formuladores de políticas. A partir daí, é essencial reconhecer o que funciona e o que não funciona e pensar em formas de aprimoramento. Embora seja fundamental questionar e compreender as dinâmicas de poder subjacentes, é preciso avançar em direção a mecanismos que empoderem a todos.

A literatura já contém parte desse trabalho desenvolvido, incluindo as principais consequências e formas de aprimoramento. No entanto, chegou o momento de criar, pensar além das limitações e juntar as peças do quebra-cabeça apresentadas neste estudo. Assim, é hora de construir mecanismos que utilizem avaliações

formativas e somativas (Brown et al., 2014) em contextos autênticos (Syverson, 2011), que promovam habilidades de ordem superior (Hayward & Spencer, 2010), que deem espaço para problemas a serem determinados (Beghetto, 2019), que gerem soluções locais e utilizem tecnologias (Behizadeh & Lynch, 2017). Além disso, a redução do alto impacto atribuído aos testes deve ser reconsiderada (Hooge et al., 2012). A introdução de recursos, como portfólios (Chudowsky & Pellegrino, 2003), projetos (BTS Project & NEPC, 2023), autoavaliação e avaliação de pares (Açıkalın, 2014), a incorporação de *softwares* que medem processos e fornecem *feedback* (Behizadeh & Lynch, 2017), entre muitas outras opções detalhadas na Figura 5, devem ser exploradas. No entanto, é necessário destacar que, seja qual for o tipo de avaliação, ela deve responder ao contexto no qual está inserida e ser um meio de promover e integrar-se ao processo de aprendizagem. Em essência, é possível estabelecer um equilíbrio entre a melhoria da aprendizagem e um mecanismo robusto para garantir a qualidade nos sistemas educacionais.

Em conclusão, as mudanças exigem que os governos tomem a iniciativa e comecem a pensar colaborativamente, fora das limitações tradicionais, para criar mecanismos de ALE. Esses mecanismos devem permitir futuros criativos e complexos, nos quais a educação preserve seu propósito principal – a aprendizagem – e permita que os alunos tenham sucesso e prosperem de acordo com seus interesses e talentos individuais (Nussbaum, 2011; Emler et al., 2019), ampliando as possibilidades da educação e buscando a equidade, em vez de perpetuar as desigualdades (Zhao, 2014).

REFERÊNCIAS

Ab Kadir, M. A. (2017). Engendering a culture of thinking in a culture of performativity: The challenge of mediating tensions in the Singaporean educational system. *Cambridge Journal of Education*, *47*(2), 227-246. https://doi.org/10.1080/0305764X.2016.1148115

Açıkalın, M. (2014). Future of social studies education in Turkey. *Journal of International Social Studies*, *4*(1), 93-102. https://www.iajiss.org/index.php/iajiss/article/view/130

Ackerman, J. (2004). Co-governance for accountability: Beyond "exit" and "voice". *World Development, 32*(3), 447-463. https://doi.org/10.1016/j.worlddev.2003.06.015

Acosta, S., Garza, T., Hsu, H.-Y., Goodson, P., Padrón, Y., Goltz, H. H., & Johnston, A. (2020). The accountability culture: A systematic review of high-stakes testing and english learners in the United States during no child left behind. *Educational Psychology Review, 32*, 327-352. https://doi.org/10.1007/s10648-019-09511-2

Anderson, K. J. (2012). Science education and test-based accountability: Reviewing their relationship and exploring implications for future policy. *Science Education*, *96*(1), 104-129. https://doi.org/10.1002/sce.20464

Ashadi, A., Margana, M., Mukminatun, S., & Utami, A. B. (2022). High stakes testing cancellation and its impact on EFL teaching and learning: Lessons from Indonesia. *International Journal of Language Education, 6*(4), 397-411. https://doi.org/10.26858/ijole.v6i4.34743

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258-267. https://doi.org/10.3102/0013189X07306523

Au, W. (2009). Social studies, social justice: W(h)ither the social studies in high-stakes testing? *Teacher Education Quarterly, 36*(1), 43-58. https://files.eric.ed.gov/fulltext/EJ851027.pdf

Bacon, J., & Pomponio, E. (2023). A call for radical over reductionist approaches to 'inclusive' reform in neoliberal times: An analysis of position statements in the United States. *International Journal of Inclusive Education, 27*(3), 354-375. https://doi.org/10.1080/13603116.2020.1858978

Ball, S. J. (2003). The teacher's soul and the terrors of performativity. *Journal of Education Policy, 18*(2), 215-228. https://doi.org/10.1080/0268093022000043065

Ball, S. J. (2012a). *Foucault, power, and education*. Routledge.

Ball, S. J. (2012b). *Global Education Inc.: New policy networks and the neoliberal imaginary*. Taylor & Francis Group.

Ball, S. J. (2012c). Performativity, commodification and commitment: An I-Spy guide to the neoliberal university. *British Journal of Educational Studies, 60*(1), 17-28. https://doi.org/10.1080/00071005.2011.650940

Ball, S. J. (2015). Education, governance and the tyranny of numbers. *Journal of Education Policy, 30*(3), 299-301. https://doi.org/10.1080/02680939.2015.1013271

Ball, S. J. (2017). *Foucault as educator*. Springer.

Beadie, N. (2019). "Hidden" governance or counterfactual case? The US failure to pass a national education act, 1870-1940. In J. Westberg, L. Boser, & I. Brühwiler (Eds.), *School acts and the rise of mass schooling: Education policy in the long nineteenth century* (pp. 325-348). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-13570-6_14

Beghetto, R. A. (2019). Large-scale assessments, personalized learning, and creativity: Paradoxes and possibilities. *ECNU Review of Education, 2*(3), 311-327. https://doi.org/10.1177/2096531119878963

Behizadeh, N., & Lynch, T. L. (2017). Righting technologies: How large-scale assessment can foster a more equitable education system. *Berkeley Review of Education, 7*(1), 25-47. https://doi.org/10.5070/B87130877

Benavot, A., Cha, Y.-K., Kamens, D., Meyer, J. W., & Wong, S.-Y. (1991). Knowledge for the masses: World models and national curricula, 1920-1986. *American Sociological Review, 56*(1), 85-100. https://doi.org/10.2307/2095675

Berliner, D. (2011). Rational responses to high stakes testing: The case of curriculum narrowing and the harm that follows. *Cambridge Journal of Education, 41*(3), 287-302. https://doi.org/10.1080/0305764X.2011.607151

Beyond Test Scores Project (BTS Project), & National Education Policy Center (NEPC). (2023, Spring). *Educational accountability 3.0: Beyond ESSA*. BTS Project; NEPC.

Boon, R., Voltz, D., Lawson, C., & Baskette, M. (2007). The impact of high-stakes testing for individuals with disabilities: A review synthesis. *Journal of the American Academy of Special Education Professionals*, 54-67. https://files.eric.ed.gov/fulltext/EJ1140225.pdf

Börzel, T. A. (2010). *Governance with/out government: False promises or flawed premises?* [Working Paper No. 23]. SFB Governance. https://ciaotest.cc.columbia.edu/wps/sfb/0018726/f_0018726_16022.pdf

Bovens, M. (2010). Two concepts of accountability: Accountability as a virtue and as a mechanism. *West European Politics, 33*(5), 946-967. https://doi.org/10.1080/01402382.2010.486119

Bovens, M., Schillemans, T., & Hart, P. T. (2008). Does public accountability work? An assessment tool. *Public Administration, 86*(1), 225-242. https://doi.org/10.1111/j.1467-9299.2008.00716.x

Brown, N. J., Afflerbach, P. P., & Croninger, R. G. (2014). Assessment of critical-analytic thinking. *Educational Psychology Review, 26*, 543-560. https://doi.org/10.1007/s10648-014-9280-4

Cato, H., & Walker, K. (2022). The influences of teacher knowledge on qualitative writing assessment. *Journal of Language and Literacy Education, 18*(2), 1-21. https://eric.ed.gov/?id=EJ1374401

Chudowsky, N., & Pellegrino, J. W. (2003). Large-scale assessments that support learning: What will it take? *Theory into Practice, 42*(1), 75-83. https://doi.org/10.1207/s15430421tip4201_10

Cimbricz, S. (2002). State-mandated testing and teachers' beliefs and practice. *Education Policy Analysis Archives, 10*(2), 1-21. https://doi.org/10.14507/epaa.v10n2.2002

Dixon-Woods, M., Cavers, D., Agarwal, S., Annandale, E., Arthur, A., Harvey, J., Hsu, R., Katbamna, S., Olsen, R., Smith, L., Riley, R., & Sutton, A. J. (2006). Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. *BMC Medical Research Methodology, 6*, Article 35. https://doi.org/10.1186/1471-2288-6-35

Domina, T., Penner, A., & Penner, E. (2017). Categorical inequality: Schools as sorting machines. *Annual Review of Sociology, 43*, 311-330. https://doi.org/10.1146/annurev-soc-060116-053354

Dorn, S. (1998). The political legacy of school accountability systems. *Education Policy Analysis Archives*, 6(1), 1-33. https://doi.org/10.14507/epaa.v6n1.1998

Ehren, M. C., Jones, K., & Perryman, J. (2016). Side effects of school inspection; motivations and contexts for strategic responses. In M. C. M. Ehren (Ed.), *Methods and modalities of effective school inspections* (pp. 87-109). Springer. https://doi.org/10.1007/978-3-319-31003-9_5

Emler, T. E., Zhao, Y., Deng, J., Yin, D., & Wang, Y. (2019). Side effects of large-scale assessments in education. *ECNU Review of Education, 2*(3), 279-296. https://doi.org/10.1177/2096531119878964

Falabella, A. (2021). The seduction of *hyper-surveillance*: Standards, testing, and accountability. *Educational Administration Quarterly, 57*(1), 113-142. https://doi.org/10.1177/0013161X20912299

Frankema, E. (2009). The expansion of mass education in twentieth century Latin America: A global comparative perspective. *Revista de Historia Económica – Journal of Iberian and Latin American Economic History, 27*(3), 359-396. https://doi.org/10.1017/S0212610900000811

Gillanders, C., Iruka, I. U., Bagwell, C., & Adejumo, T. (2021). Parents' perceptions of a K-3 formative assessment. *School Community Journal, 31*(2), 239-266. https://files.eric.ed.gov/fulltext/EJ1323055.pdf

Gough, D., Oliver, S., & Thomas, J. (2013). *Learning from research: Systematic reviews for informing policy decisions*. Alliance for Useful.

Gough, D., Oliver, S., & Thomas, J. (2017). Introducing systematic reviews. In D. Gough, S. Oliver, & J. Thomas (Eds.), *An introduction to systematic reviews* (2nd ed., pp. 1-17). Sage.

Gough, D., Thomas, J., & Oliver, S. (2012). Clarifying differences between review designs and methods. *Systematic Reviews, 1*, Article 28. https://doi.org/10.1186/2046-4053-1-28

Green, A. (2013). *Education and state formation: Europe, East Asia and the USA*. Palgrave Macmillan.

Gregory, K., & Clarke, M. (2003). High-stakes assessment in England and Singapore. *Theory into Practice, 42*(1), 66-74. https://doi.org/10.1207/s15430421tip4201_9

Hamilton, L. S., Schwartz, H. L., Stecher, B. M., & Steele, J. L. (2013). Improving accountability through expanded measures of performance. *Journal of Educational Administration, 51*(4), 453-475. https://doi.org/10.1108/09578231311325659

Harlen, W., & Crick, R. D. (2003). Testing and motivation for learning. *Assessment in Education: Principles, Policy & Practice, 10*(2), 169-207. https://doi.org/10.1080/0969594032000121270

Harlen, W., Crick, R. D., Broadfoot, P., Daugherty, R., Gardner, J., James, M., & Stobart, G. (2002). *A systematic review of the impact of summative assessment and tests on students' motivation for learning* (EPPI-Centre Review). EPPI-Centre.

Hayward, L. (2015). Assessment is learning: The preposition vanishes. *Assessment in Education: Principles, Policy & Practice*, *22*(1), 27-43. https://doi.org/10.1080/0969594X.2014.984656

Hayward, L., Priestley, M., & Young, M. (2004). Ruffling the calm of the ocean floor: Merging practice, policy and research in assessment in Scotland. *Oxford Review of Education*, *30*(3), 397-415. https://doi.org/10.1080/0305498042000260502

Hayward, L., & Spencer, E. (2010). The complexities of change: Formative assessment in Scotland. *Curriculum Journal*, *21*(2), 161-177. https://doi.org/10.1080/09585176.2010.480827

Helfenbein, R. J. (2004). New times, new stakes: Moments of transit, accountability, and classroom practice. *Review of Education, Pedagogy, and Cultural Studies*, *26*(2-3), 91-109. https://doi.org/10.1080/10714410490480368

Herman, J. L., & Winters, L. (1994). Portfolio research: A slim collection. *Educational Leadership*, *52*(2), 48-55.

Hood, C., & Dixon, R. (2016). Not what it said on the tin? Reflections on three decades of UK public management reform. *Financial Accountability & Management*, *32*(4), 409-428. https://doi.org/10.1111/faam.12095

Hooge, E., Burns, T., & Wilkoszewski, H. (2012). *Looking beyond the numbers: Stakeholders and multiple school accountability* [Working Papers No. 85]. OECD Education. https://dx.doi.org/10.1787/5k91dl7ct6q6-en

Hopewell, S., Clarke, M., & Mallett, S. (2005). Grey literature and systematic reviews. In H. R. Rothstein, A. J. Sutton, & M. Borenstein, *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 49-72). Wiley. https://doi.org/10.1002/0470870168.ch4

Hoyuelos, A., & Cabanellas, I. (1996). Malaguzzi y el valor de lo cotidiano [Presentación de trabajo]. *Congreso de Pamplona*, Pamplona, España. https://www.waece.org/biblioteca/pdfs/d091.pdf

Huddleston, A. P., & Rockwell, E. C. (2015). Assessment for the masses: A historical critique of high-stakes testing in reading. *Texas Journal of Literacy Education*, *3*(1), 38-49. https://eric.ed.gov/?id=EJ1110955

Hutchinson, C., & Hayward, L. (2005). The journey so far: Assessment for learning in Scotland. *Curriculum Journal*, *16*(2), 225-248. https://doi.org/10.1080/09585170500136184

Jones, M. G., & Ennes, M. (2018). High-stakes testing. *Oxford Bibliographies*. https://doi.org/10.1093/obo/9780199756810-0200

Kamens, D. H., & McNeely, C. L. (2010). Globalization and the growth of international educational testing and national assessment. *Comparative Education Review*, *54*(1), 5-25. https://doi.org/10.1086/648471

Lee, J. (2008). Is test-driven external accountability effective? Synthesizing the evidence from cross-state causal-comparative and correlational studies. *Review of Educational Research*, *78*(3), 608-644. http://www.jstor.org/stable/40071139

Levi-Faur, D. (2012). From "big government" to "big governance"? In D. Levi-Faur (Ed.), *The Oxford Handbook of Governance* (pp. 3-18). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199560530.013.0001

Lingard, B. (2009). Testing times: The need for new intelligent accountabilities for schooling. *QTU Professional Magazine*, *24*, 13-19.

Link, A. N., & Scott, J. T. (2010). Historical perspectives on public accountability. In A. N. Link, & J. T. Scott, *Public goods, public gains: Calculating the social benefits of public R &C D* (pp. 20-26). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199729685.003.0003

Lynn, L. E. (2012). The many faces of governance: Adaptation? Transformation? Both? Neither? In D. Levi-Faur (Ed.), *The Oxford Handbook of Governance* (pp. 49-64). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199560530.013.0004

Madaus, G., & Russell, M. (2010). Paradoxes of high-stakes testing. *Journal of Education, 190*(1-2), 21-30. https://doi.org/10.1177/0022057410190001-205

Moss, G. (2022). Researching the prospects for change that COVID disruption has brought to high stakes testing and accountability systems. *Education Policy Analysis Archives, 30*(139), 1-24. https://doi.org/10.14507/epaa.30.6320

Muller, J. Z. (2018). *The tyranny of metrics*. Princeton University Press.

Murphy, M. (2021). *Social theory: A new introduction*. Palgrave Macmillan.

Nichols, S. L. (2007). High-stakes testing: Does it increase achievement? *Journal of Applied School Psychology, 23*(2), 47-64. https://doi.org/10.1300/J370v23n02_04

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Harvard Education Press.

Nussbaum, M. C. (2011). *Creating capabilities: The human development approach*. Harvard University Press.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Wiley-Blackwell.

Ramirez, F. O., & Boli, J. (1987). The political construction of mass schooling: European origins and worldwide institutionalization. *Sociology of Education, 60*(1), 2-17. https://doi.org/10.2307/2112615

Rhodes, R. A. W. (1996). The new governance: Governing without government. *Political Studies, 44*(4), 652-667. https://doi.org/10.1111/j.1467-9248.1996.tb01747.x

Rhodes, R. A. W. (2007). Understanding governance: Ten years on. *Organization Studies, 28*(8), 1243-1264. https://doi.org/10.1177/0170840607076586

Roberson, S. (2011). Defying the default culture and creating a culture of possibility. *Education, 131*(4), 885-904.

Schillemans, T. (2016). Calibrating public sector accountability: Translating experimental findings to public sector accountability. *Public Management Review, 18*(9), 1400-1420. https://doi.org/10.1080/14719037.2015.1112423

Schillemans, T., Van Twist, M., & Vanhommerig, I. (2013). Innovations in accountability: Learning through interactive, dynamic, and citizen-initiated forms of accountability. *Public Performance & Management Review, 36*(3), 407-435. https://doi.org/10.2753/PMR1530-9576360302

Sigvardsson, A. (2017). Teaching poetry reading in secondary education: Findings from a systematic literature review. *Scandinavian Journal of Educational Research, 61*(5), 584-599. https://doi.org/10.1080/00313831.2016.1172503

Soysal, Y. N., & Strang, D. (1989). Construction of the first mass education systems in nineteenth-century Europe. *Sociology of Education, 62*(4), 277-288. https://doi.org/10.2307/2112831

Suto, I., & Oates, T. (2021). *High-stakes testing after basic secondary education: How and why is it done in high-performing education systems?* (Research report). Cambridge Assessment.

Syverson, M. A. (2011). Social justice and evidence-based assessment with the learning record. In P. Kriese, & R. E. Osborne (Eds.), *Social justice, poverty and race: Normative and empirical points of view* (pp. 93-102). Brill.

University College London (UCL). (2023). Systematic reviews: Stages in a systematic review. *UCL*. Retrieved May 10, 2023 from https://library-guides.ucl.ac.uk/systematic-reviews/stages

Verger, A., Fontdevila, C., & Parcerisa, L. (2019). Reforming governance through policy instruments: How and to what extent standards, tests and accountability in education spread worldwide. *Discourse: Studies in the Cultural Politics of Education*, *40*(2), 248-270. https://doi.org/10.1080/01596306.2019.1569882

Verger, A., Parcerisa, L., & Fontdevila, C. (2019). The growth and spread of large-scale assessments and test-based accountabilities: A political sociology of global education reforms. *Educational Review*, *71*(1), 5-30. https://doi.org/10.1080/00131911.2019.1522045

Volante, L. (2007). Educational quality and accountability in Ontario: Past, present, and future. *Canadian Journal of Educational Administration and Policy*, (58), 1-21. https://journalhosting.ucalgary.ca/index.php/cjeap/article/view/42739

Wang, Y. (2017). The social networks and paradoxes of the opt-out movement amid the Common Core State Standards implementation: The case of New York. *Education Policy Analysis Archives*, *25*(34), 1-27. https://doi.org/10.14507/epaa.25.2757

Westberg, J., Boser, L., & Brühwiler, I. (2019). The history of school acts. In J. Westberg, L. Boser, & I. Brühwiler (Eds.), *School acts and the rise of mass schooling: Education policy in the long nineteenth century* (pp. 1-15). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-13570-6_1

Wyse, D., Hayward, L., & Pandya, J. Z. (Eds.). (2015). *The Sage handbook of curriculum, pedagogy and assessment*. Sage.

Zhao, Y. (2014). *Who's afraid of the big bad dragon? Why China has the best (and worst) education system in the world*. John Wiley & Sons.

Zimmerman, B. J., & Dibenedetto, M. K. (2008). Mastery learning and assessment: Implications for students and teachers in an era of high-stakes testing. *Psychology in the Schools*, *45*(3), 206-216. https://doi.org/10.1002/pits.20291

Zinkina, J., Korotayev, A., & Andreev, A. (2016). Mass primary education in the nineteenth century. In L. E. Grinin, I. V. Ilyin, P. Herrmann, & A. V. Korotayev (Eds.), *Globalistics and globalization studies: Global transformations and global future* (pp. 63-70). 'Uchitel' Publishing House.

APÊNDICES

Apêndice A – Palavras-chave sub-revisão 1

- S1 – [testes de alto impacto OU testes de alto impacto em escolas]
- S2 – revisão sistemática OU revisão sistemática da literatura OU revisão da literatura
- S3 – S1 E S2: [testes de alto impacto OU testes de alto impacto em escolas] E [revisão sistemática OU revisão sistemática da literatura OU revisão da literatura]

Mais os filtros: Periódicos acadêmicos (revisados por pares).

Apêndice B – Palavras-chave sub-revisão 2

Pesquisa final: **S21** E **S22**

- **S21** continha: **S19** E **S20**
- **S19** continha os seguintes termos:
- AB [ambiente de SALA DE AULA OU inovações EDUCACIONAIS OU EDUCAÇÃO & estado OU avaliação FORMATIVA OU estratégias de APRENDIZAGEM OU programas EDUCACIONAIS OU avaliação OU ESTUDANTES OU classificação de OU educação no ENSINO MÉDIO OU educação PRÉ-ESCOLAR OU educação BÁSICA OU ESCÓCIA OU VALORIZAÇÃO da ciência POLÍTICA OU planejamento CURRICULAR OU mudança EDUCACIONAL OU avaliação FORMATIVA OU programas EDUCACIONAIS OU avaliação ESCÓCIA avaliação FORMATIVA OU desenvolvimento do PROFESSOR; educação SECUNDÁRIA; educação BÁSICA OU ESCÓCIA OU mudança ORGANIZACIONAL OU inovações EDUCACIONAIS OU PESQUISA OU testes & medidas EDUCACIONAIS OU educação SECUNDÁRIA OU educação BÁSICA OU avaliação EDUCACIONAL OU PESQUISA OU padrões EDUCACIONAIS DOS PROFESSORES OU pesquisa sobre o ambiente ESCOLAR OU ativismo dos ESTUDANTES OU movimentos de PROTESTO] OU educação alternativa AB OU avaliação alternativa AB OU avaliação alternativa AB E testes de alto impacto OU avaliação em larga escala E resistência OU mudança OU novas formas OU novas possibilidades OU [novas possibilidades na avaliação educacional OU avaliação em larga escala]
- **S20** continha os seguintes termos:
  testes de alto impacto AB OU avaliação em larga escala AB E avaliação AB OU testes padronizados AB OU testes padronizados AB OU performatividade AB
- **S22** continha os seguintes termos:
  inovações EDUCACIONAIS OU avaliação formativa OU mudança EDUCACIONAL OU avaliação para aprendizagem

Apêndice C – Detalhes dos especialistas consultados na Revisão Sistemática da Literatura

Especialista número 1: Clive Dimmock

Recomendou Louise Hayward:

1. Hutchinson, C., & Hayward, L. (2005). The journey so far: Assessment for learning in Scotland. *Curriculum Journal, 16*(2), 225-248. https://doi.org/10.1080/09585170500136184

2. Hayward, L., Priestley, M., & Young, M. (2004). Ruffling the calm of the ocean floor: Merging practice, policy and research in assessment in Scotland. *Oxford Review of Education, 30*(3), 397-415. https://doi.org/10.1080/0305498042000260502
3. Hayward, L., & Spencer, E. (2010). The complexities of change: Formative assessment in Scotland. *Curriculum Journal, 21*(2), 161-177. https://doi.org/10.1080/09585176.2010.480827

Especialista número 2: Clara Fontdevila

Recomendou os seguintes textos e autores:

1. Hooge, E., Burns, T., & Wilkoszewski, H. (2012). *Looking beyond the numbers: Stakeholders and multiple school accountability* [Working Papers No. 85]. OECD Education. https://dx.doi.org/10.1787/5k91dl7ct6q6-en
2. Lingard, B. (2009). Testing times: The need for new intelligent accountabilities for schooling. *QTU Professional Magazine, 24*, 13-19.
3. Beyond Test Scores Project (BTS Project), & National Education Policy Center (NEPC). (2023, Spring). *Educational accountability 3.0: Beyond ESSA*. BTS Project; NEPC.
4. Domina, T., Penner, A., & Penner, E. (2017). Categorical inequality: Schools as sorting machines. *Annual Review of Sociology, 43*, 311-330. https://doi.org/10.1146/annurev-soc-060116-053354