A psicologia da avaliação: abordando as pressões que os professores enfrentam

Gavin Thomas Lumsden Brown[1]

Resumo

Os processos de avaliação são usados para informar formalmente melhorias e, sumariamente, para determinar o valor do ensino, das escolas e dos alunos. Esses usos criam pressões sobre os professores porque eles têm objetivos conflitantes em relação à melhoria da aprendizagem dos estudantes e responsabilização dos docentes. Este artigo apresenta uma visão geral da compreensão do autor sobre a avaliação e os desafios que a responsabilização apresenta às concepções que os professores têm sobre a avaliação, e identifica como seu maior obstáculo a reduzida informação sobre as avaliações educacionais. Uma solução para este dilema é apresentada ao descrever o *software Assessment tools for teaching and learning* (em português, Ferramentas de avaliação para ensino e aprendizagem), utilizado em escolas da Nova Zelândia.

Palavras-chave: Psicologia dos professores; Testes educacionais; Elaboração de relatórios educacionais.

Abstract

The psychology of evaluation: Addressing the pressures teachers face

Evaluation processes are used to formatively inform improved instruction and summatively to determine the worth or value of teaching, schools, and students. These uses create pressures for teachers because they have conflicting goals around improvement vs. accountability. This perspective paper overviews the author's understanding of evaluation, the challenges accountability presents to the conceptions teachers have about assessment and identifies a major obstacle to the impact of evaluation in the low information level of most so-called educational tests. A solution to the dilemma is recommended in the example of the Assessment Tools for Teaching and Learning software deployed in New Zealand schools.

Keywords: Teacher psychology; educational testing; design of test reports.

Resumen

La psicología de la evaluación: abordar las presiones que enfrentan los docentes

Los procesos de evaluación se utilizan para informar formalmente las mejoras y, brevemente, para determinar el valor de la enseñanza, las escuelas y los estudiantes. Estos usos crean presiones sobre los maestros porque tienen objetivos contradictorios con respecto a mejorar el aprendizaje de los estudiantes y la responsabilidad de los maestros. Este artículo presenta una visión general de la comprensión del autor sobre la evaluación y los desafíos que presenta la rendición de cuentas para

[1] University of Auckland, Auckland, Nova Zelândia.

las concepciones de los docentes sobre la evaluación, e identifica la información reducida sobre las evaluaciones educativas como su mayor obstáculo. Se presenta una solución a este dilema describiendo las herramientas de evaluación para el software de enseñanza y aprendizaje, que se utilizan en las escuelas de Nueva Zelanda.
Palabras Clave: Psicología de los docentes; Pruebas educativas; Elaboración de informes educativos.

Introdução

Em muitas sociedades, os testes educacionais são usados em alunos para fins de privilégios, como conclusão de curso e bolsas de estudos. Na última metade do século XX, as respostas dos alunos em testes padronizados para avaliar a qualidade da escola e dos professores eram utilizadas, acarretando graves consequências negativas. Ao mesmo tempo, espera-se que os professores monitorem a aprendizagem dos alunos e a eficácia do seu ensino, para avaliar o progresso e o desempenho dos estudantes, antes mesmo das métricas terem consequências substanciais para o aluno e para o professor.

A tensão relacionada aos objetivos da avaliação cria desafios e pressões significativas para os professores. Neste artigo quero primeiro descrever o que engloba a análise de métricas na educação e como ela se relaciona com a avaliação e os testes. Quero esclarecer as pressões que a avaliação de rendimento gera para os educadores. A minha análise me leva a focar nas insuficiências dos testes utilizados na educação em relação à incapacidade de oferecer *insights* para a melhoria do trabalho de educadores. Em sequência, analiso um sistema de avaliação usado na Nova Zelândia que mostra como os testes alinhados ao currículo, padronizados e assistidos por computador podem ajudar os professores a cumprirem a função formativa da avaliação, satisfazendo as questões de responsabilidade dos administradores. No modelo descrito, os professores são tratados como profissionais que precisam de ferramentas mais eficientes.

Compreendendo a avaliação

O termo "avaliação" está atrelado, no dicionário do Instituto de Ciências da Educação do Departamento de Educação dos Estados Unidos (Educational Research Information Center [ERIC], 2001), à "qualificação". Esse é o termo mais utilizado, inclusive fora do mundo anglófono, para se referir a análises qualitativas da educação. Assim, a ideia de avaliação está dentro da qualificação. Como consequência, o termo "qualificação" refere-se a todos e quaisquer meios de reunir informações sobre o

desempenho ou necessidades/pontos fortes e as interpretações ou usos subsequentes aos quais esses dados são colocados (Brown, 2018).

Os métodos de avaliação incluem análise em escolas de salas de aula e processos formais, como testes, provas, portfólios (inclusive digitais), observações, análise de desempenho, avaliação por pares ou autoavaliação e assim por diante. O termo também abrange concursos públicos, qualificações usadas para determinar a entrada em programas competitivos ou níveis de escolaridade e a concessão de marcos sociais importantes, como a graduação.

Qualificar significa medir a qualidade (ou seja, os pontos fortes/fracos) dos produtos ou processos de trabalho, mas também emitir julgamentos sobre o valor, mérito ou significado. Possibilita respostas se os produtos ou os processos atendem às expectativas, se devem ser aprovados ou reprovados, se são satisfatórios ou excelentes. As avaliações são necessárias para informar as decisões se um produto ou processo é ou não é bom o suficiente, se precisa de mudanças e quais seriam elas (Brown, 2018). Isto nos leva à pergunta principal: qual é o objetivo da avaliação?

Há uma perspectiva tecnológica (Cheung, 2000), derivada da avaliação que é usada na engenharia, que fornece informações úteis sobre o que é necessário na educação. Ao planejar um voo tripulado para a Lua, os engenheiros precisam ter como objetivo levar os humanos à Lua e trazê-los de volta com segurança (ou seja, o Projeto Apollo da década de 1960). Todas as avaliações ao longo do processo tiveram que ser alinhadas a este objetivo para que fossem válidas. Ao aplicar um "teste", ele precisa fornecer informações sobre se e como o trabalho realizado até o momento está atrelado ao objetivo (ou seja, é provável que cheguemos à Lua com o que temos?). Seja qual for a resposta, as avaliações devem também fornecer informações sobre o que deve ser feito para voltar ao caminho correto e/ou permanecer nele. À medida que a cápsula viaja para a Lua, as decisões são tomadas com base em interpretações sobre os cenários avaliados e suas possibilidades.

Além disso, os engenheiros precisam saber com antecedência qual será a reação de cada ação. O resultado de cada escolha, baseada em testes desenvolvidos tecnologicamente, é esperado – não é uma incógnita. Os engenheiros têm confiança na previsibilidade de suas ações. Claro que o caso da astrofísica é mais simples do que no ensino; mas os professores precisam da avaliação para cumprir objetivos semelhantes. Os professores precisam: conhecer seus objetivos curriculares, de avaliações que

informem sobre como as turmas e os estudantes estão, de um conhecimento sólido sobre o que devem fazer em resposta aos resultados da avaliação e as consequências previsíveis de suas ações.

Formativa e somativa

Como Scriven (1967) deixou claro, existem dois momentos-chave nos quais as avaliações podem ser executadas. Se as avaliações são realizadas em qualquer momento antes das avaliações finais e com objetivo de informar potenciais mudanças ao processo e melhorar o resultado, são denominadas avaliações formativas. A ideia-chave é verificar se é preciso fazer modificações antes que seja tarde demais. As avaliações formativas coletam dados sobre a aprendizagem dos alunos para informar o planejamento dos professores e as atividades de ensino, para garantir que os alunos atinjam seus objetivos de aprendizagem, compreensão e ação (Bloom, Hastings, & Madaus, 1971). A qualidade das avaliações formativas deve ser alta para informar corretamente aos professores, administradores e alunos o que precisa ser modificado. Caso contrário, as decisões que deveriam promover a melhoria podem estar inadequadas (Hattie & Brown, 2010).

O segundo momento de avaliação, ao final de um período de ensino, é chamado de somativo. É uma descrição e uma avaliação qualitativa do que está sendo avaliado, sem que haja possibilidade de modificar seu resultado. É aqui que a ideia de classificação revela a noção-chave de "valor", implicando quantificação e mérito. Neste momento, é comum que a qualidade geral seja o foco (por exemplo, levando em consideração o desempenho geral do aluno, conclui-se que seu trabalho foi satisfatório, não bom, portanto, recebe nota média). Considerando a consequência da avaliação somativa, a qualidade da avaliação deve ser elevada, pois, caso contrário, as decisões avaliativas estarão equivocadas.

As avaliações formativas deveriam ser tão bem-feitas que não haveria surpresas nas decisões somativas a serem tomadas. Além disso, como Stake (ver Scriven, 1991) deixou claro, as avaliações somativas podem tornar-se formativas se os utilizadores do teste optarem por explorar as informações obtidas. Há países que utilizaram os seus resultados do Programa Internacional de Avaliação de Estudantes (Pisa), em conjunto com outros fatores, para fazer mudanças substanciais e benéficas nas políticas e práticas (Teltemann & Klieme, 2016). Assim, a diferença fundamental entre as avaliações formativas e as avaliações somativas, segundo Scriven, é no calendário e não no formato ou nas características de qualidade.

Isso contrasta com a visão criada por Sadler (1989), na qual o autor afirma que as avaliações formativas e somativas possuem formatos distintos. Ao associar fins somativos a provas e exames formais, Sadler procurou remover interpretações arriscadas do processo de melhoria da qualidade da aprendizagem dos alunos ou da instrução dos professores. Mas, ao fazê-lo, ele removeu os testes formais das ferramentas de avaliação formativa. Esta tendência continuou na avaliação *para* a aprendizagem (Black & Wiliam, 1998), movimento em que a avaliação formativa se tornou um conjunto de interações informais entre aluno e instrutor e, simultaneamente, entre alunos (Leahy, Lyon, Thompson, & Wiliam, 2005). Esta visão pedagógica da avaliação (Stobart, 2006) centrou-se no que os alunos faziam na avaliação, reduzindo consideravelmente o papel desempenhado pelo professor. Mais recentemente, especialmente no ensino superior (Yan & Boud, 2022), em que se espera que os alunos, como adultos, sejam responsáveis por seus próprios resultados, os teóricos colocam a avaliação *como* aprendizagem; uma postura que parece inseparável da autorregulação da aprendizagem (Brown, 2022). Os leitores interessados podem consultar Brown (2021) para uma visão integrada de como essas diferentes abordagens de avaliação se desenrolam na formação de professores.

Múltiplos propósitos de avaliação

A visão da avaliação promulgada na Nova Zelândia (Brown, 2004; 2006a; 2008) pode ser resumida em quatro propósitos. Sucintamente, a avaliação pode ser formativa, mas precisa fornecer diferentes tipos de informações às duas partes da educação (ou seja, educador e aluno). Os alunos precisam de informações sobre como estão se saindo, o que estão fazendo bem ou mal e em que precisam focar seu próprio estudo. Os educadores, por outro lado, precisam saber não apenas a situação de cada aluno, mas também como estão as turmas, quais são as necessidades prioritárias, o que os alunos dominaram e no que já não exigem tanta instrução e quem dentro do grupo têm necessidades semelhantes. Naturalmente, este uso orientado para a melhoria da avaliação educacional depende de as avaliações serem indicadores válidos e precisos do conhecimento, habilidade e desenvolvimento dos alunos.

Paralelamente a esta utilização formativa, encontram-se avaliações somativas a serem feitas, que oferecem diferentes ferramentas para alunos e educadores. Classificar com precisão o desempenho dos alunos (por exemplo, baixo, médio, alto) permite a validação de decisões futuras, como quem se forma, quem recebe uma bolsa de estudos

ou quem precisa repetir um ano de escolaridade. Do mesmo modo, o desempenho dos alunos nas avaliações somativas pode ser utilizado para inferir a qualidade das escolas e dos professores. De fato, em muitos locais, a qualidade do ensino é avaliada através de testes aplicados com alunos. Por exemplo, a forma como os governos e os meios de comunicação tratam os resultados dos sistemas internacionais de avaliação em larga escala (Pisa ou *Trends in International Mathematics and Science Study* – TIMSS) que são vistos como uma forma para avaliar a qualidade de todo o sistema educacional do país. Nesse sentido, a qualidade geral dos esforços curriculares pode ser avaliada em sistemas de monitoramento (por exemplo, o *National Assessment of Educational Progress* – Naep, dos Estados Unidos da América – EUA, ou o *National Monitoring Study of Student Achievement* – NMSSA, da Nova Zelândia) que utilizam amostras de estudantes para inferir o que funciona ou não. Outros locais implementam censos sistemáticos de todos os alunos (por exemplo, testes *Key Stage*, do Reino Unido) para determinar se as escolas estão garantindo que todos os alunos atendam às expectativas curriculares. Assim, através de avaliações somativas, tanto os alunos como as escolas/professores podem ser responsabilizados pelos recursos investidos e por seu desempenho em relação aos parâmetros estabelecidos.

Em resposta às avaliações somativas, alguns educadores têm uma perspectiva que denominei "irrelevante". Eles veem as avaliações como sendo ruins para os alunos (por exemplo, rótulos de "reprovação"), injustas com aqueles de origens carentes que não podem atingir o domínio do currículo no ritmo recomendado e que são desnecessárias. Esses professores veem os atos formais e os usos da avaliação como algo que podem ignorar, porque já sabem o que os alunos entendem, fazem e precisam. Eles aplicam as avaliações por serem uma exigência do seu trabalho, mas não colocam esforços em tais atos, porque conhecem e cuidam de seus alunos.

Vale ressaltar que Remesal (2011) também capta essa tensão entre o foco em professores x alunos e em técnicas formativas x somativas. Consequentemente, a política e a prática de avaliação devem resolver as tensões entre os usos formativos e os somativos, seja em relação ao desempenho de professores e alunos, seja para fornecer informações para professores e alunos. Independentemente destas decisões políticas, é importante considerar a forma como as decisões de avaliação têm impacto nos participantes. Não é de surpreender que o participante-chave seja o professor, dado que é o responsável pela gestão das atividades da sala de aula e pelo currículo.

O desafio da accountability

Infelizmente, para aqueles que não gostam de atribuir mérito, é uma verdade inconveniente que, como sociedade, precisamos de avaliações (descritivas e avaliativas). As avaliações têm de ser robustas, medindo o que realmente valorizamos. Por exemplo, a qualidade de uma redação não é determinada pela precisão da ortografia, pontuação ou gramática, embora sejam as mais fáceis de identificar. Avaliações robustas medem a habilidade, o conhecimento e o desempenho tão bem que podemos ter confiança nas decisões vinculadas a elas. Essa concepção engloba validade por confiabilidade das informações (Cizek, 2020).

Isso significa que avaliações robustas acertam mais do que erram. O custo de aprovação daqueles que não dominam o esperado acarreta danos ao bem comum. Isso significa que precisamos de avaliações robustas para assegurar à sociedade que aqueles que se saíram suficientemente bem podem receber privilégios (carteira de motorista, certificação como engenheiro, médico, contador etc.) enquanto aqueles que não o fizeram são impedidos de exercer os direitos e privilégios relativos ao sucesso. Como Lingard e Lewis (2016, p. 400) apontam:

> A oposição total à responsabilização não é uma posição justificável. Em vez disso, precisamos, progressivamente, rever os conceitos da responsabilização que reconheça os propósitos societais mais amplos da escolaridade e que responsabilize os sistemas, as escolas e as comunidades, tanto de baixo para cima como de cima para baixo.

Aplicando esta abordagem à avaliação educacional, Resnick e Resnick (1989) expuseram um desenho denominado "O que você testa é o que você tem" (em inglês, "*what you test is what you get*" abreviado pelos autores como WYTIWYG). Em WYTIWYG, parte-se do pressuposto de que, se algo é importante, deve ser avaliado, caso contrário, será visto como opcional e pode não ser ensinado ou aprendido. Se não estiver em uma avaliação formal, a sociedade não pode ter certeza de que os principais conhecimentos ou habilidades estão sendo trabalhados. Argumentaram ainda que a importância desse conhecimento pode ser desconsiderada se não houver consequências com os testes. As consequências geralmente anunciadas incluíam a mudança de escolas, professores ou líderes (por exemplo, substituir ou demitir professores ou fechar a escola). A aplicação desta lógica nos testes de responsabilização nos EUA revelou resultados indesejáveis. A revisão realizada por Nichols e Harris (2016) mostrou que testar os alunos para avaliar a qualidade do professor é uma ideia contraproducente. De um modo geral, esta abordagem da responsabilização leva à distorção do ensino,

do currículo e do profissionalismo dos professores, oprime os estudantes das minorias, entedia os estudantes altamente capacitados e incentiva a fraude nesses testes.

Este resultado está totalmente em consonância com a revisão de Lerner e Tetlock (1999) sobre os efeitos conhecidos da responsabilização. O primeiro efeito é a entrega de resultados tal como foi especificado. Quando os superiores querem pontuações mais altas nos testes e melhores taxas de aprovação, os professores fazem o possível para obtê-los. Infelizmente, isso pode levar a um comportamento antiético ou não profissional "justificado" por parte dos professores e escolas para gerenciar a forma como seu trabalho é visto por seus superiores. Por exemplo, há mais de 30 anos, Cannell (1989) documentou um desempenho acima da média em testes padronizados em cada estado dos EUA por uma série de razões, incluindo: (1) os professores ensinaram aos alunos diretamente o que caía no teste, (2) as escolas usaram testes antigos que tinham normas de comparação inválidas, (3) os professores tiveram acesso a testes antigos e planejaram as aulas nesse sentido, (4) as escolas incentivaram os alunos de baixo desempenho a não irem à escola nos dias de teste, (5) os professores deram dicas aos alunos durante o teste e (6) as escolas corrigiram as respostas dos testes antes de enviá-las à agência regulatória central. Este último processo resultou na prisão de professores e líderes escolares em Atlanta, EUA, por falsificar o desempenho dos alunos nos testes de desempenho escolar[2].

As fortes consequências negativas para as pontuações baixas parecem ser vistas por muitos professores como fundamentalmente injustas, em parte porque o sistema escolar não é geralmente justo. Por exemplo, algumas escolas têm maior quantidade de alunos cujos pais vivem na pobreza ou não tiveram formação adequada. Outras escolas são pobres, carecem de recursos essenciais ou encontram-se em locais onde é difícil atrair professores altamente qualificados. Psicologicamente, a maioria dos professores se opõe a rotular os alunos como fracassados ou a usar testes padronizados para avaliar uma criança. Nesse sentido, "trapacear" para aumentar os resultados parece ser uma atitude racional, ainda que pouco profissional. Em um sistema de responsabilização, que ignora barreiras à performance adequada de alunos nos testes, há mais probabilidade de ações de resistência e subversão e defesa destas.

Isto significa que é um exercício frágil tentar entender o desempenho do sistema escolar por meio de alunos, para julgar professores e escolas. Como Bonner (2016) aponta elegantemente, a intenção de fazer um bom trabalho e ver os alunos

[2] https://en.wikipedia.org/wiki/Atlanta_Public_Schools_cheating_scandal

aprenderem é pressionada por mandatos políticos, políticas e distribuição de recursos. Assim, antes de culparmos os professores por quererem evitar "más" notícias de que o seu ensino nem sempre funciona, temos que examinar a forma como os professores concebem e utilizam a avaliação. Pode ser que os professores não sejam culpados; pelo contrário, pode ser que a natureza e a qualidade dos recursos de avaliação fornecidos aos professores não apoiem o objetivo diagnóstico e formativo (Brown & Hattie, 2012).

Concepções de avaliação dos professores

Com base nesta análise de quatro conceitos para fins de avaliação, o Registro de Concepções de Avaliação para Professores (em inglês, *Teachers Conceptions of Assessment Inventory*, [Brown, 2006b]) foi desenvolvido na Nova Zelândia com professores do Ensino Fundamental. Como a Nova Zelândia tinha um sistema de ensino primário fortemente centrado na criança, sem testes ou exames formais, não foi surpreendente notar que os professores eram fortemente a favor da avaliação para aprendizagem, rejeitaram a ideia de que a avaliação era irrelevante, concordaram cautelosamente que a realização de testes poderia avaliar os alunos e rejeitaram o pensamento de que a avaliação poderia identificar a qualidade da escola (Brown, 2004).

O apoio à ideia de que as avaliações trazem benefícios à aprendizagem foi encontrado em diversos contextos geográficos e níveis de ensino. Professores primários e secundários na Nova Zelândia (Brown, 2011), Hong Kong (Brown, Kennedy, Fok, Chan, & Yu, 2009), Queensland (Brown, Lake, & Matters, 2011), Chipre (Brown & Michaelides, 2011), Equador (Brown & Remesal, 2017) e Egito (Gebril & Brown, 2014) foram os que mais aprovaram a ideia de avaliação para melhora na aprendizagem. Nesses mesmos locais, os professores rejeitam veementemente o uso da avaliação para a responsabilização da escola (Nova Zelândia, Egito e Queensland) ou a ideia de que a avaliação é irrelevante (Hong Kong, Equador e Chipre). Os professores na Índia apoiam a avaliação para melhora da aprendizagem e avaliação diagnóstica (Brown, Chaudhry, & Dhamija, 2015). É importante notar que, entre os professores, a prioridade da avaliação para melhoria da aprendizagem é perceptível, independentemente da cultura ou da consequência associada à avaliação.

Verificou-se também que diferentes concepções de avaliação influenciam os tipos que são utilizadas pelos professores (Brown, 2009). A escolha por formatos de avaliação informais, em oposição a testes formais controlados pelo professor, foi

influenciada por concepções como a da avaliação como irrelevante e sem relação com a melhoria do aprendizado. Este resultado sugeriu que os professores tratavam as avaliações formais como irrelevantes e consideravam que as avaliações informais forneciam informações úteis sobre a aprendizagem. Ao mesmo tempo, professores acreditavam que os testes formais funcionavam como medição superficial do ganho de conhecimento dos alunos. Em contraste, surge a percepção de que uma maior compreensão da qualidade das escolas depende de avaliações profundas. Juntos, estes indicadores mostram que os professores da Nova Zelândia pensam que os testes desenvolvidos externamente para os alunos mediam a aquisição de conhecimentos e o foco de uma escola de qualidade é incentivar um pensamento profundo, relacional e crítico.

De um modo geral, os professores consideram que a avaliação pode contribuir para a melhoria do ensino e da aprendizagem. Rejeitam a ideia de que é irrelevante. Eles estão cientes de que essas avaliações qualificam os alunos e têm sentimentos mistos sobre isso. Há indícios de que as crenças dos professores não interferem nos resultados dos alunos nas avaliações formais e padronizadas. Parece simples fazer recomendações de desenvolvimento profissional dos professores para melhorar a compreensão sobre a avaliação, mas o desafio é o sistema de avaliação e de teste.

Então, o que há de errado na avaliação?

A fonte de informação mais difundida para avaliação consiste em testes e provas. Possuem a virtude de serem padronizados para todos os participantes e por exigirem alguns pressupostos. O primeiro é que os resultados dos testes são "medidas não contaminadas, presumindo que as respostas de um aluno aos itens de teste refletem apenas seus conhecimentos e habilidades no conteúdo alvo" (Wise & Smith, 2016, p. 207). O segundo, como disse Thorndike (1924), "todas as nossas medições assumem que o indivíduo em questão se esforça tanto quanto pode para obter a melhor pontuação possível" (p. 228). Portanto, quaisquer falhas ou lapsos momentâneos de atenção, motivação ou pensamentos são apenas parte do contexto de "erro". A partir disso, os pais presumem que os resultados dos testes dizem a verdade sobre o que os alunos sabem e o quanto trabalharam antes e durante um teste (Harris & Brown, 2016). Não é difícil concluir que esses pressupostos são falhos.

Testes e provas raramente vão além de uma pontuação geral, normalmente uma porcentagem de acertos. Uma vez que temos uma pontuação é fácil criar uma

classificação atrelada (por exemplo, o 1º da turma ou o último). Testes padronizados muitas vezes classificam o desempenho em uma pontuação transformada (por exemplo, a escala de QI tem mediano = 100 e desvio-padrão = 15). Testes padronizados tendem a relatar a ordem de classificação em relação a uma amostra normativa usando métodos como porcentagem ou escala de cinco pontos. Estas pontuações não fornecem informações diagnósticas ou formativas sólidas, mesmo se fornecidas no início do ano letivo ou durante o período de ensino.

Um dos motivos para a ausência dessas informações se deve ao fato da abordagem clássica avaliar cada item com o mesmo peso. Além de ser um cálculo simplista de adição dos acertos, há outras razões pelas quais os resultados podem ser inflados. Testes com muitas questões ou perguntas fáceis produzem pontuações altas, com muitos alunos ficando acima da média (por exemplo, a nota de corte para aprovação). Isso pode aparentar bom desempenho do professor e levar os alunos a acreditarem que estão caminhando bem e dentro do esperado. Se não tivermos pontuações ajustadas pela dificuldade, como ocorre na teoria de resposta ao item (TRI), professores e alunos serão seduzidos pela complacência se pontuações altas forem obtidas em testes fáceis. Só é possível saber quem dominou apenas as questões fáceis e quem pode fazer as mais difíceis quando os testes trazem questões desafiadoras. Consequentemente, saber pontuar ponderando a dificuldade do item significa que é possível identificar questões fáceis que só precisam de prática versus as difíceis que precisam de instrução e quais alunos estão em cada condição.

As ordens de classificação são fáceis de determinar e entender em um ambiente de sala de aula. Apenas três podem subir ao pódio como vencedores da prova. No entanto, a ordem não diz ao professor, ao aluno ou aos pais o que cada aluno precisa aprender em seguida. Um efeito colateral infeliz da ordem de classificação é que ela leva à suposição de que aqueles que estão na base não podem aprender enquanto aqueles que estão no topo não podem aprender ainda mais. Ambas as conclusões são falsas e antieducativas. Relacionado a isso, os professores tendem a usar seus próprios grupos como métrica, presumindo que o melhor e o pior desempenho em suas turmas são iguais ao melhor ou pior do país. É aqui que os testes padronizados podem ajudar, comparando os alunos a partir de uma população nacionalmente representativa. Dessa forma, podemos dizer coisas como "este rapaz é o pior da turma, mas tem um desempenho igual ao terço superior do país. Seus problemas não são graves".

Outra razão para indicar que a pontuação total é educacionalmente ineficaz é o fato de o diagnóstico ter origem nos componentes do currículo que os professores precisam ensinar. Nenhum professor ensina apenas um assunto; eles planejam lições sequencialmente ordenadas que se baseiam em material previamente ensinado. Isso significa que, ao avaliar a nossa instrução, precisamos saber quais partes do currículo os alunos já compreendem e quais precisamos começar a ensinar ou, eventualmente, revisar. Se não tivermos perguntas pensadas em relação ao currículo, seus objetivos, vertentes ou componentes que os professores devem ensinar, as avaliações apenas reforçarão o que os professores já sabem. Os professores tendem a notar rápido quem é o melhor e o pior em uma turma, mas não sabem "quem precisa aprender o quê". É nesse aspecto que testes bem elaborados e formulados para obter os objetivos de ensino podem ajudar (Brown & Hattie, 2012).

Um bom sistema educacional fornece aos seus professores ferramentas que podem fornecer informações úteis em termos educativos. Testes diagnósticos e padronizados podem informar aos professores:

- Quais são os objetivos que os alunos dominaram e não precisam mais ser ensinados;
- Quais são os objetivos que tiveram bons resultados e pelos quais podem ser elogiados;
- Quais são os objetivos que devem rever, revisar e praticar; e
- Quais são os objetivos que os estudantes ainda precisam de instrução.

No entanto, projetar e analisar testes dessa maneira em tempo hábil requer habilidades e recursos que a maioria dos professores não possui. Apesar dos muitos livros didáticos e cursos que ensinam o desenvolvimento e a análise de testes, a experiência mostra que a maioria dos professores não tem as competências necessárias para elaborar testes de alta qualidade. Eles também não têm tempo ou prioridade para analisar e compreender os resultados dos testes. Embora possa ser tentador desconsiderar os testes formais como uma ferramenta para os professores, nenhum outro mecanismo permite que os professores descubram de forma tão eficiente quem sabe o quê e quem precisa aprender o quê. Além disso, como a pesquisa sobre as crenças dos professores mostrou, elas não são realmente o problema. O verdadeiro problema parece estar nos testes educativos inadequados e mal concebidos.

A resposta da Nova Zelândia: asTTle

Para enfrentar este desafio, um sistema de avaliação tecnológica foi desenvolvido há 20 anos na Nova Zelândia e continua a ser gerido pelas escolas. O sistema Instrumentos de Avaliação do Ensino e da aprendizagem (*Assessment Tools for Teaching and Learning*, e-asTTle[3]) fornece testes padronizados para compreensão de leitura, matemática e escrita em duas línguas (inglês e te reo Māori, a língua indígena de Aotearoa). Com um financiamento substancial do governo, desenvolvemos na Universidade de Auckland uma plataforma de testes em computador e controlada pela escola. Ela evoluiu gradualmente de computadores de redes locais para uma plataforma *web* com testes adaptáveis por computador.

Alinhamento ao currículo

O que torna o sistema poderoso não são as suas tecnologias informáticas ou estatísticas (Hattie & Brown, 2008). Em vez disso, os princípios fundamentais dos testes educacionais foram implantados no projeto desde o início. Isso inclui vincular todas as tarefas, perguntas e relatórios aos objetivos, vertentes e níveis de progressão do currículo. O sistema permite aos professores e às escolas escolherem quando testam, quem testam, o seu conteúdo e como utilizam os dados. Para conseguir isso, o sistema possibilita que os professores especifiquem quais áreas curriculares e níveis de dificuldade devem ser contidos em um teste criado a partir do banco de perguntas pré-calibradas. Os professores podem visualizar o teste antes de administrá-lo para garantir que ele se adapta às suas intenções e ao ensino realizado em sala de aula. Se o mecanismo de criação do teste não produzir o que o professor tinha em mente, ele pode ser alterado no conteúdo e na dificuldade.

Todas as informações sobre o desempenho são claramente comunicadas após extensas análises psicométricas, utilizando a TRI, para dar um peso maior a questões mais exigentes. Com base em dados de 30.000 alunos em leitura, 25.000 em matemática e 20.000 em escrita, são fornecidas normas para os anos escolares 4-12 e para os níveis curriculares 2-6. A dificuldade de um item foi derivada da informação normativa usando métodos da TRI (Embretson & Reise, 2000). Itens que poucos alunos acertaram têm níveis de dificuldade mais altos do que aqueles que a maioria dos alunos acertou. Mas, em vez de apenas dar um número, as pontuações de desempenho foram mapeadas para

[3] https://e-asttle.tki.org.nz/

os níveis curriculares usando painéis de professores. Dessa forma, um professor sabe onde na progressão curricular uma criança ou classe está e se eles estão à frente ou atrás das normas relevantes do ano. Isso possibilita a análise de cada um e a comparação com os outros; um conjunto útil de percepções para compartilhar com os pais e famílias. Os usuários podem ter confiança no alinhamento com os padrões curriculares e as normas de desempenho usadas para criar relatórios para seus próprios alunos.

Para apoiar o desafio de avaliar o currículo, o sistema asTTle faz o trabalho repetitivo de entrada de dados, pontuação de itens e agregação de itens curriculares significativos, deixando o professor livre para pensar sobre como projetar e fornecer instruções apropriadas em vez de tentar descobrir "quem atingiu o quê no teste". Para garantir às escolas que essas avaliações eram realmente para fins diagnósticos e formativos, elas mantinham o controle sobre o sistema e seus dados. Isso significava que nenhuma autoridade central recolhia os dados, podia examiná-los ou dizer quem teria que ser testado. Isso significava que as escolas e os professores podiam ver se havia "más notícias" em seus próprios dados e responder a elas bem antes de qualquer avaliador externo chegar à escola (Hattie & Brown, 2008).

Visualização de dados

A maior conquista da asTTle reside nos nossos esforços para comunicar aos professores as informações que precisam para tomarem decisões educacionais sólidas e embasadas sobre o seu ensino. Uma revisão da literatura deixou claro que a maioria das visualizações de dados em torno da avaliação são mal compreendidas pelos leitores pretendidos (Brown, 2001). A nossa busca por bons exemplos encontrou o CRESST *Quality School Portfolio* (Baker, 1999), que utilizou um sistema de relatórios baseado em sinais de trânsito (ou seja, vermelho = alerta, abaixo da média; amarelo = média; verde = bom, acima da média) para comunicar informações e ajudar os professores a compreender os dados que as escolas possuem. Foram coletados dados sobre as informações que os professores queriam após os testes (Meagher-Lundberg, 2000). Com a ajuda de um *designer* gráfico, os modelos foram construídos com os professores (Meagher-Lundberg, 2001ª; 2001b), o que resultou em projetos utilizados desde o início do processo. (Hattie, Brown, & Keegan, 2003).

O sistema gera um conjunto de relatórios que permitem aos professores e dirigentes escolares avaliar, responder e monitorar os efeitos do seu trabalho.

Cada relatório é gerado para informar cada membro da escola (ou seja, professor, gestor de departamento, diretor) e, por isso, as informações são apresentadas de forma a facilitar o uso de seus resultados. As especificidades dos relatórios e a sua lógica subjacente são apresentadas em Brown, O'Leary e Hattie (2018). Os relatórios visualizam os pontos fortes e fracos do desempenho dos alunos nos níveis individual e de grupo, de modo que possam ser usados pelos professores para planejar a instrução, relatar aos pais e com os pais e dar retorno aos alunos (Archer & Brown, 2013).

O desempenho geral de um estudante foi determinado pelo peso do TRI dos itens respondidos corretamente, sem qualquer penalidade por adivinhação ou erro. Isso significa que o desempenho dos alunos pode ser classificado pelos itens respondidos corretamente e sua dificuldade. Seguindo o Método Wright de exibição (*Wright Map display method* [Wright & Stone, 1979]), a Trajetória de Aprendizagem Individual (em inglês, *Individual Learning Pathways*, ou ILP) posicionou os itens e objetivos do desempenho de um aluno em um dos quadrantes de um gráfico. Os números dos itens são apresentados entre parênteses ao lado do seu objetivo para dar prioridade ao currículo ao invés das perguntas do teste. Cada quadrante conduz a mensagens de ensino e *feedbacks* adequados, tal como indicado no Quadro. Este relatório tem sido utilizado com sucesso nas escolas como base das discussões entre pais e professores sobre o comportamento de uma criança e os passos seguintes.

Quadro – Guia interpretativo dos caminhos de aprendizagem individual

Cor	Conteúdo	Interpretação
Verde	Correto, mas fácil	Este material foi dominado, então o professor não deve mais ensiná-lo.
Amarelo	Correto, mas difícil	Parabéns! Isso mostra o que você pode alcançar. O professor deve garantir mais atividades deste nível e tipo.
Vermelho	Incorreto, mas fácil	O aluno deve ser encorajado a praticar mais, o professor deve verificar se foi ensinado, mas sem centralizar, já que será fácil para o aluno corrigir.
Azul	Incorreto, mas difícil	O professor deve planejar como ensinar porque foi feito de forma errada, sugerindo que requer conhecimento ou habilidade que ainda não foram ensinados. É improvável que seja aprendido por osmose ou métodos de descoberta.

Ao contrário da profissão médica que trata um paciente de cada vez, o professor de sala de aula lida com grupos de 20 a 40 alunos na Nova Zelândia. Em outros países esse grupo pode ser maior. Além disso, em muitos sistemas escolares, os dirigentes de nível médio (de séries ou disciplinas) querem ver os pontos fortes e fracos em uma coorte de um determinado grupo. Assim, desenvolvemos um relatório agregado que identifica a proporção do grupo para cada objetivo de realização, utilizando o mesmo esquema de cores codificado do ILP. Para efeitos de planejamento, este relatório mostra a dimensão do Azul em todo o grupo em relação aos objetivos. A interpretação é semelhante à do ILP; os professores devem focar no ensino da zona azul e minimizar a zona verde. Relatórios indicam que os departamentos escolares que fazem isso veem um crescimento robusto no desempenho dos alunos; em outras palavras, os alunos aprendem o que lhes ensinamos e a avaliação pode mostrar isso.

Na Nova Zelândia, pelo menos, agrupamento é uma prática comum dentro das turmas (Wilkinson & Townsend, 2000), bem como a tutoria entre pares (Roscoe & Chi, 2007). Para atender a esses objetivos, foi desenvolvido um gráfico de distribuição dos grupos em cada nível curricular, dando ao professor uma referência visual da disparidade no desempenho. Ao clicar no gráfico, uma lista de nomes é gerada em cada subnível curricular, permitindo aos professores a capacidade de ver quem poderia ser agrupado para instrução diferenciada (Moon, 2016) ou as parcerias na tutoria entre pares.

Em vez de apenas comparar as normas nacionais, os diretores escolares podem selecionar e comparar escolas com características demográficas semelhantes (por exemplo, "Escolas como a minha" [Hattie, 2002]). Desta forma, é feita uma comparação mais fidedigna. Se outras escolas como a minha têm, em média, desempenho melhor do que a minha escola, já não posso dizer que é por causa da minha comunidade. Desta forma, as escolas são desafiadas a deixar de culpar fatores externos e, em vez disso, considerar o que e como as outras escolas estão fazendo para melhorar o desempenho.

Aceitação

Para proteger o potencial formativo da asTTle, ficou claro ao longo do seu desenvolvimento, tanto para o Ministério da Educação como para a comunidade escolar, que o sistema continha as perguntas e as respostas. Isso significava que seria possível "trapacear" para fazer uma turma ou escola parecer boa. No entanto, não haveria valor

educativo em fazê-lo, porque somente quando os educadores olham para os alunos não sabem ou são capazes de fazer que os professores podem começar a trabalhar para enfrentar esses desafios.

Outros mecanismos que apoiaram a aceitabilidade incluem o desenvolvimento gradual e a transparência. O sistema asTTle desenvolveu-se através de múltiplas variações tecnológicas de acordo com o sistema de infraestrutura escolar até se modificar para ser *on-line*, como uma aplicação *web* (Brown, 2019). O sistema asTTle reconheceu a contribuição dos milhares de professores que estiveram envolvidos como palestrantes, avaliadores de itens e revisores; um passo que ajudou a ganhar aceitabilidade em todo o país. Além disso, foram publicados relatórios técnicos[4], dando transparência ao projeto asTTle, o que foi feito, o que havíamos aprendido e à razão pela qual havíamos feito certas escolhas. Esta transparência ajudou os professores a aceitar o valor do sistema como instrumento formativo.

Conclusão

Nesta abordagem da avaliação formativa, os testes não são o problema. Os testes mal elaborados e com relatórios limitados são o problema. O sistema na Nova Zelândia possibilitou a criação de um sistema de testes educacionalmente eficaz que ajude as escolas e os professores a fazerem o seu trabalho, mas também satisfazem as expectativas de responsabilização. Isso pode ser feito porque os professores e as suas atitudes não estão no centro do problema. Os professores querem ensinar melhor, mas muitas vezes são deixados à deriva com sistemas de avaliação tão limitados que não podem ajudar os professores a identificar o que precisa ser ensino e a quem. Encarar a busca de professores por melhorias no ensino como vantagem, em vez de problema, foi um grande passo. Um possível *slogan* seria "se você respeita os professores e quer melhorar os resultados, dê a eles uma ferramenta como asTTle". Isso é válido porque, em vez de se tratar de tecnologia ou pontuação somativa, o sistema é concebido como uma tecnologia educacional que ajuda os professores a fazer o que precisam fazer.

Qualquer sociedade que pretenda melhorar a aprendizagem dos alunos deve criar um sistema de avaliação que ultrapasse o *status quo*. O sistema deve estar centrado em dizer aos professores o que eles precisam para melhorar as suas atividades de

[4] https://e-asttle.tki.org.nz/Reports-and-research/asTTle-technical-reports

ensino em sala de aula. As avaliações devem ser estreitamente integradas ao currículo, de modo a realmente servirem aos objetivos de ensino. O exemplo asTTle na Nova Zelândia é um exemplo bem-sucedido da utilização de práticas avaliativas aliadas à tecnologia para melhorar o ensino.

Referências

Archer, E., & Brown, G. T. L. (2013). Beyond rhetoric: leveraging learning from New Zealand's assessment tools for teaching and learning for South Africa. *Education as Change*, *17*(1), 131-147. https://doi.org/10.1080/16823206.2013.773932

Baker, E. L. (1999, jul.). Technology: something's coming-something good. *Cresst Policy Brief*, 2. Recuperado em 23 de setembro de 2022 de https://cresst.org/publications/cresst-publication-3245/

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Bloom, B., Hastings, J., & Madaus, G. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw Hill.

Bonner, S. M. (2016). Teachers' perceptions about assessment: competing narratives. In G. T. L. Brown, & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 21-39). London: Routledge.

Brown, G. T. (2019). Technologies and infrastructure: costs and obstacles in developing large-scale computer-based testing. *Education Inquiry*, *10*(1), 4-20. https://doi.org/10.1080/20004508.2018.1529528

Brown, G. T. L. (2001, Aug.). Reporting assessment information to teachers: report of project asTTle outputs design. *Technical Report 15: Reporting Assessment Information.* Recuperado em 23 de setembro de 2022 de http://tinyurl.com/q69ew9a

Brown, G. T. L. (2006b). *Teachers' conceptions of assessment inventory-Abridged (TCoA-IIIA-Version 3-Abridged)*. Auckland: University of Auckland.

Brown, G. T. L. (2004). Teachers' conceptions of assessment: implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 301-318. https://doi.org/10.1080/0969594042000304609

Brown, G. T. L. (2006a). Teachers' conceptions of assessment: validation of an abridged version. *Psychological Reports*, *99*(1), 166-170. https://doi.org/10.2466/pr0.99.1.166-170

Brown, G. T. L. (2008). *Conceptions of assessment: understanding what assessment means to teachers and students*. New York: Nova Science.

Brown, G. T. L. (2009). Teachers' self-reported assessment practices and conceptions: using structural equation modelling to examine measurement and structural models. In T. Teo & M. S. Khine (Eds.), *Structural equation modelling in educational research: concepts and applications* (pp. 243-266). Rotterdam: Sense.

Brown, G. T. L. (2011). Teachers' conceptions of assessment: comparing primary and secondary teachers in New Zealand. *Assessment Matters*, *3*, 45-70. https://doi.org/10.18296/am.0097

Brown, G. T. L. (2018). *Assessment of student achievement*. London: Routledge.

Brown, G. T. L. (2021). Evaluation in teacher training: an analysis of theoretical conceptions. In W. Santos, & R. Stieg (Eds.), *Evaluación educativa en la formación de profesores: Brasil, Colombia, Chile, España, Inglaterra, México, Nueva Zelanda y Uruguay* (pp. 53-67). Curitiba: Appris.

Brown, G. T. L. (2022). Assessments cause and contribute to learning: if only we let them. In Z. Yan, & L. Yan (Eds.), *Assessment as learning: maximising opportunities for student learning and achievement* (pp. 38-52). London: Routledge.

Brown, G. T. L., & Michaelides, M. P. (2011). Ecological rationality in teachers' conceptions of assessment across samples from Cyprus and New Zealand. *European Journal of Psychology of Education*, *26*(3), 319-337. https://doi.org/10.1007/s10212-010-0052-3

Brown, G. T. L., & Remesal, A. (2017). Teachers' conceptions of assessment: comparing two inventories with Ecuadorian teachers. *Studies in Educational Evaluation*, *55*, 68-74. https://doi.org/10.1016/j.stueduc.2017.07.003

Brown, G. T. L., Chaudhry, H., & Dhamija, R. (2015). The impact of an assessment policy upon teachers' self-reported assessment beliefs and practices: a quasi-experimental study of Indian teachers in private schools. *International Journal of Educational Research*, *71*, 50-64. https://doi.org/10.1016/j.ijer.2015.03.001

Brown, G. T. L., Kennedy, K. J., Fok, P. K., Chan, J. K. S., & Yu, W. M. (2009). Assessment for student improvement: understanding Hong Kong teachers' conceptions and practices of assessment. *Assessment in Education: Principles, Policy & Practice, 16*(3), 347-363. https://doi.org/10.1080/09695940903319737

Brown, G. T. L., Lake, R., & Matters, G. (2011). Queensland teachers' conceptions of assessment: The impact of policy priorities on teacher attitudes. *Teaching and Teacher Education, 27*(1), 210-220. https://doi.org/10.1016/j.tate.2010.08.003

Brown, G. T. L., O'Leary, T. M., & Hattie, J. A. C. (2018). Effective reporting for formative assessment: the asTTle case example. In D. Zapata-Rivera (Ed.), *Score reporting: research and applications* (pp. 107-125). London: Routledge.

Brown, G. T., & Hattie, J. (2012). The benefits of regular standardized assessment in childhood education: guiding improved instruction and learning. In S. Suggate, & E. Reese (Eds.), *Contemporary educational debates in childhood education and development* (pp. 287-292). London: Routledge.

Cannell, J. J. (1989). *How public educators cheat on standardized achievement tests.* Wheaton: Friends for Education.

Cheung, D. (2000). Measuring teachers' meta-orientations to curriculum: application of hierarchical confirmatory analysis. *Journal of Experimental Education, 68*(2), 149-165. https://doi.org/10.1080/00220970009598500

Cizek, G. J. (2020). *Validity: an integrated approach to test score meaning and use.* London: Routledge.

Educational Research Information Center – ERIC. (2001). *Thesaurus of ERIC descriptors* (14th ed.). Phoenix: Oryx.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah: Lawrence Erlbaum Associates.

Gebril, A., & Brown, G. T. L. (2014). The effect of high-stakes examination systems on teacher beliefs: egyptian teachers' conceptions of assessment. *Assessment in Education: Principles, Policy & Practice, 21*(1), 16-33. https://doi.org/10.1080/0969594X.2013.831030

Harris, L. R., & Brown, G. T. L. (2016). Assessment and parents. In M. A. Peters (Ed.), *Encyclopedia of educational philosophy and theory* (pp. 1-6). Berlin: Springer.

Hattie, J. A. (2002). Schools like mine: cluster analysis of New Zealand schools. *Technical Report 14, Project asTTle.* Recuperado em 23 de setembro de 2022 de https://e-asttle.tki.org.nz/content/download/1470/5943/version/1/file/14.+Cluster+analysis+of+NZ+schools+2002.pdf

Hattie, J. A. C., Brown, G. T. L., & Keegan, P. J. (2003). A national teacher-managed, curriculum-based assessment system: assessment tools for teaching & learning (asTTle). *International Journal of Learning, 10*, 771-778.

Hattie, J. A., & Brown, G. T. (2010). Assessment and evaluation. In C. Rubie-Davies (Ed.), *Educational psychology: concepts, research and challenges* (pp. 116-131). London: Routledge.

Hattie, J. A., & Brown, G. T. L. (2008). Technology for school-based assessment and assessment for learning: development principles from New Zealand. *Journal of Educational Technology Systems, 36*(2), 189-201. https://doi.org/10.2190/ET.36.2.g

Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment minute by minute, day by day. *Educational Leadership, 63*(3), 18-24.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125*(2), 255-275. https://doi.org/10.1037/0033-2909.125.2.255ra

Lingard, B., & Lewis, S. (2016). Globalization of the anglo-american approach to top-down, test-based educational accountability. In G. T. L. Brown, & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 387-403). London: Routledge.

Meagher-Lundberg, P. (2000, out.). Comparison variables useful to teachers in analysing assessment results. *Technical Report 1: Project asTTle.* Recuperado em 23 de setembro de 2022 de https://e-asttle.tki.org.nz/content/download/1454/5895/version/1/file/1.+Comparison+Variables+2000.pdf

Meagher-Lundberg, P. (2001a). Output reporting design: focus group 1. *Technical Report 9: Project asTTle.* Recuperado em 23 de setembro de 2022 de https://e-asttle.tki.org.nz/content/download/1464/5925/version/1/file/9.+Output+reporting+design+Focus+Group+1+2001.pdf

Meagher-Lundberg, P. (2001b). Output reporting design: focus group 2. *Technical Report 10: Output Focus Group 2.* Recuperado em 23 de setembro de 2022 de

https://e-asttle.tki.org.nz/content/download/1465/5928/version/1/file/10.+Output+reporting+design+Focus+Group+2+2001.pdf

Moon, T. R. (2016). Differentiated instruction and assessment: an approach to classroom assessment in conditions of student diversity. In G. T. L. Brown, & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 284-301). London: Routledge.

Nichols, S. L., & Harris, L. R. (2016). Accountability assessment's effects on teachers and schools. In G. T. L. Brown, & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 40-56). London: Routledge.

Remesal, A. (2011). Primary and secondary teachers' conceptions of assessment: a qualitative study. *Teaching and Teacher Education, 27*(2), 472-482. https://doi.org/10.1016/j.tate.2010.09.017

Roscoe, R. D., & Chi, M. T. H. (2007). Understanding tutor learning: knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research, 77*(4), 534-574. https://doi.org/10.3102/0034654307309920

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119-144. https://doi.org/10.1007/BF00117714

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (Vol. 1, pp. 39-83). Chicago: Rand McNally.

Scriven, M. (1991). Beyond formative and summative evaluation. In M. W. McLaughlin, & D. C. Phillips (Eds.), *Evaluation & education: at quarter century* (Vol. 90, part II, pp. 19-64). Chicago: National Society for the Study of Education.

Stobart, G. (2006). The validity of formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 133-146). Thousand Oaks: Sage.

Teltemann, J., & Klieme, E. (2016). The impact of international testing projects on policy and practice. In G. T. L. Brown, & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 369-386). London: Routledge.

Wilkinson, I. A. G., & Townsend, M. A. R. (2000). From Rata to Rimu: grouping for instruction in best practice New Zealand classrooms. *The Reading Teacher, 53*(6), 460-471.

Wise, S. L., & Smith, L. F. (2016). The validity of assessment when students don't give good effort. In G. T. L. Brown, & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 204-220). London: Routledge.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* San Diego: Mesa.

Yan, Z., & Boud, D. (2022). Conceptualising assessment-as-learning. In Z. Yan, & L. Yang (Eds.), *Assessment as learning: maximising opportunities for student learning and achievement* (pp. 11–24). London: Routledge.

Submetido em: setembro de 2022
Aceito em: outubro de 2022

Sobre o autor

Gavin Thomas Lumsden Brown
Doutor em Psicologia da Educação pela University of Auckland. Professor da Faculdade de Educação e Trabalho Social da University of Auckland, Nova Zelândia. E-mail: gt.brown@auckland.ac.nz