TYPE Perspective
PUBLISHED 17 July 2023
DOI 10.3389/feduc.2023.1211580

Challenges and opportunities in score reporting: a panel of personal perspectives

Gavin T. L. Brown[1]*, Priya Kannan[2], Sandip Sinharay[3], Diego Zapata-Rivera[3] and April L. Zenisky[4]

[1]Faculty of Education and Social Work, The University of Auckland, Auckland, New Zealand, [2]WestEd, San Francisco, CA, United States, [3]Educational Testing Service, Princeton, NJ, United States, [4]College of Education, University of Massachusetts, Amherst, MA, United States

OPEN ACCESS

EDITED BY
Xinya Liang,
University of Arkansas, United States

REVIEWED BY
Alexander Robitzsch,
IPN - Leibniz Institute for Science and Mathematics Education, Germany
Yi-Fang Wu,
Cambium Assessment, Inc.,
United States

*CORRESPONDENCE
Gavin T. L. Brown
✉ gt.brown@auckland.ac.nz

RECEIVED 24 April 2023
ACCEPTED 28 June 2023
PUBLISHED 17 July 2023

CITATION
Brown GTL, Kannan P, Sinharay S, Zapata-Rivera D and Zenisky AL (2023) Challenges and opportunities in score reporting: a panel of personal perspectives. *Front. Educ.* 8:1211580.
doi: 10.3389/feduc.2023.1211580

The field of score reporting continues to evolve because of new challenges, opportunities, and needs of society (e.g., COVID, remote teaching and learning). In this paper, the new challenges and opportunities in score reporting are discussed from the personal perspective of four experts who have previously conducted research in designing score reports in education. Comments are organized around four key questions concerning challenges raised by the Covid pandemic, how research will change, what current research is being conducted, and new directions in the field of score reporting.

KEYWORDS

post-COVID technology solutions, score reporting systems, testing, scientific communication, stakeholder comprehension, data-driven decision-making, assessment uses/purposes, data visualization

Introduction

It has been more than 4 years since the publication of *Score Reporting Research and Applications* (Zapata-Rivera et al., 2018). This book, part of the National Council for Measurement in Education (NCME) book series, includes work in areas such as validity in score reporting (Tannenbaum, 2018), cognitive affordances of graphical representations (Hegarty, 2018), evaluation of subscores (Sinharay et al., 2018), communicating measurement error information to teachers and parents (Zapata-Rivera et al., 2018), score reporting issues for licensure, certification, and admissions programs (O'Donnell and Sireci, 2018), communicating growth (Zenisky et al., 2018), score reports for large-scale testing programs (Slater et al., 2018), and evaluating the use of interactive reports and dashboards in formative contexts (Brown et al., 2018; Corrin, 2018; Feng et al., 2018).

In 2023, Diego Zapata-Rivera organized a panel of experts in NCME to discuss new challenges and opportunities in score reporting that respond to new trends in assessment due to changes in society and education. For example, we can see an increase in the use of digital learning and assessment systems which has resulted in the field of score reporting moving toward general reporting systems that provide teachers and learners with relevant insights based on an abundance of process and response data.

Four authors from the NCME book were contributors to the conference panel and are co-authors of this manuscript. They are: Gavin T. L. Brown (GTLB; the University of Auckland), Priya Kannan (PK; WestEd), April Zenisky (AZ; University of Massachusetts, Amherst), and Sandip Sinharay (SS; Educational Testing Service). Unfortunately, Linda Corrin (Deakin University), who had planned to participate in the panel discussion, was not able to participate. This manuscript captures what each panelist said in response to four questions:

Q1. How have the challenges of the last few years (e.g., COVID, remote teaching and learning) impacted the field of score reporting?
Q2. How will the nature of research in the area of score reporting change due to the availability of data from digital learning and assessment systems?
Q3. What aspects of your work can inform current work on designing and evaluating interactive reports and dashboards?
Q4. What new challenges and opportunities you expect will be present in the field of score reporting?

The order in which panelists answered questions was changed after each question give each of the panelists the opportunity to provide the first response to a question. The responses provided by the panelists are presented below followed by some final concluding remarks.

> Q1. How have the challenges of the last few years (e.g., COVID, remote teaching and learning) impacted the field of score reporting?

PK: The COVID-19 pandemic resulted in unprecedented disruptions to the ways in which kids experienced school (Parks et al., 2021; Krause et al., 2022). The drastic shift to virtual learning created learning setbacks for almost all children (Wyse et al., 2020), but particularly for children who were already underserved (Bailey et al., 2021; Goudeau et al., 2021), leading to disadvantages, both educationally and economically, that could last a lifetime (Dorn et al., 2020). The ways in which children across the various demographic and socio-economic subgroups experienced remote learning was dramatically different. With that, the learning loss varied significantly by access to remote learning, the quality of remote instruction, home support, and the degree of engagement (Dorn et al., 2020). Researchers hypothesized that the achievement gaps would start further widening (Bailey et al., 2021), and this became abundantly clear in the United States with the release of the 2022 NAEP report card.[1] There were greater score declines in reading and mathematics, particularly for Black and Hispanic students when compared to their White peers (Sparks, 2022). All of this has led to a reckoning of sorts in the score reporting community, particularly in how we look at student performance and achievement, specifically the dramatic and disparate learning setbacks experienced by students from under-served communities, and how we could shift the focus from achievement gap to opportunity gap by appropriately highlighting underlying systemic issues through our reporting.

Scholars have been discussing the impact of the expansive "opportunity gap" that exists across racial and associated class lines for several decades now (e.g., McDonnell, 1995; Porter, 1995). For example, in Ladson-Billings, 2006 AERA presidential address, Gloria Ladson-Billings highlighted how the focus on the "gap" is misplaced and called out the importance of paying attention to the "educational debt" that has accumulated over years of systemic disparities that have disproportionately impacted Black and Brown communities. However, the awareness and movement to really shift the narrative of how student test results are communicated, from a focus on 'achievement gap' to a focus on the 'opportunity gaps' for students across diverse communities (more precisely, gaps in their 'opportunities to learn') has really happened post COVID.

> To quote Marion (2020),
>
> "Opportunity-to-learn (OTL) is a more than 50-year-old concept that has evolved from a focus on whether students have had sufficient access to instruction or content linked to particular concepts, to a more robust conception regarding the conditions and resources provided to schools to enable students to succeed." (p. 2)

The key point being made is that OTL encompasses a much broader reflection on the set of resources that influence teaching and learning in the school setting. This includes factors such as school and classroom conditions, school climate, access to qualified teachers, opportunities for teacher professional development (PD) available across various districts, time scheduled for instruction across different districts, schools, and communities, student opportunities for out-of-school learning, student access to high-quality books in and out of school, student access to technological tools, and other resources.

The headlines emphasizing the performance/achievement gaps among student subgroups were persistent prior to the pandemic, but have been particularly so with the release of the most recent NAEP Report card (e.g., Mahnken, 2022; Modan, 2022). However, taking an OTL approach to score reporting and presenting high-quality OTL data alongside achievement results can help avoid the stereotyping of lower-performing groups, by pointing to some of the systemic and resource factors that influence performance. So, there is clearly a need to shift away from reporting average scores across disaggregated subgroups, which has the unintended consequence of reinforcing implicit racial, cultural, and economic stereotypes and deficit notions about groups of students, and move toward a focus on more systemic issues. We could begin to do this by using data from a variety of contextual OTL factors as the primary disaggregating variables and using effective visualization to present the within group differences among students from various demographic subgroups with different opportunities to learn (e.g., how do students from various demographic subgroups with similar access to instructional resources compare?). With such a view, we could push the needle in changing the narrative from a story that leads to deficit notions to a story that points to more systemic issues that need to be addressed. The hope is that this would then be the first step in moving toward a conversation around educational equity.

In the USA, NAEP score reports have already started considering these issues by incorporating their survey questionnaire data that includes several such OTL factors to see how these contextual variables can help explain the within group differences among student groups, and also to help policymakers identify the systemic opportunity gaps and address them appropriately. However, this shift away from focusing on achievement gaps and starting to unravel the systemic gaps in OTL should percolate beyond just NAEP reporting to score reporting on state summative assessments and beyond.

AZ: In reflecting back on the challenges of the past few years, one thing that has become abundantly clearer is the need to be realistic about test type and test purpose. It has long been the case that good data (and the communication of good data) was critically important in educational testing (Goodman and Hambleton, 2004). But, the reporting for summative tests in the context of K-12 education has

[1] https://nces.ed.gov/nationsreportcard/

historically served more of a confirmatory/passive/high-level informative purpose, because it generally takes a long time to return results and thus when the reports come, they are far removed from instructional utility (e.g., the kids are moved up a grade, have different teachers, changed schools; Hattie and Brown, 2008; Brown et al., 2018).

So, while typical summative tests may play a role moving forward, the need for data - good data, in a timely manner - is what matters more than ever. End-of-year tests may serve a policy purpose, but what is needed now are tests with direct classroom relevance. In terms of reporting, that means the priority is probably not grand, highly stable scores and performance classifications, but data for instruction that is closely linked to what a teacher can do in the next day, the next week, or the next month. So, for those of us who work in reporting, the challenges of the past few years have reinforced that we need to reorient ourselves a little.

- First, we should not try to make summative *Individual Student Reports* (ISRs) into something they cannot be. Where those kinds of tests are administered and used going forward, the reports can and should be oriented to fulfill a descriptive or informational purpose relative to the intended users, without trying to extract deeply diagnostic information where is does not exist.
- And second, in terms of reporting, we should be paying a different kind of attention to different tests (interim or through-year tests, or formative assessments) in terms of how we can craft reporting resources for the user groups articulated for those tests that target and accomplish different goals, such as instructional planning at a sufficiently usable grain size (O'Donnell and Zenisky, 2020). The results of these assessments need to be immediately and obviously connected to a lesson or activities or next steps. With reporting, the greater the distance to action, the less likely *any* action is going to occur.

All of this is not to say that we should discard summative test reporting, but in the years to come I would like to see more acknowledgment of this reality, that some tests provide information that is backward-looking, and others are built to provide information that is forward-looking. It would be helpful to stop thinking that all tests can be all things, in terms of reporting, start recognizing the different purposes of assessments and their data, and play to the strengths of each type of test as a data source and different intended users and use cases.

GTLB: The covid pandemic required education online, including assessment of student learning. Ensuring the security of online testing (Dawson, 2021) is a *sine qua non* of valid reporting. The current state of online proctoring creates doubts as to the validity of invigilation (Alessio et al., 2017; Wuthisatian, 2020). Despite efforts to ensure academic integrity, dishonesty is widespread (Murdock et al., 2016). In the context of distance examination systems, there was more cheating than previously (Reisenwitz, 2020). Further, changes made to administration and scoring of tests and examinations during the pandemic to accommodate the lockdown did not always go to plan as seen in the UK's school exam grading controversy in which an algorithm created very different grades than estimated by schoolteachers[2]. Consequently, much research is needed on how to design and validate test reports that correctly identify and communicate the impact of construct irrelevant factors on the estimation of ability or proficiency.

Governments constantly review and revise curriculum to reflect evolving perspectives of what children need to be taught, and by implication what tests need to measure. Changes in curriculum require changes to test reporting interfaces. Underneath a change of report labels lies the question of whether Curriculum Label A really means the same thing as Label Q and if the items do measure Q when designed for A. In Brown et al. (2018), there is a report for achievement objectives in reading comprehension (Figure 8.3) showing results for finding information, knowledge, understanding, and connections derived from the New Zealand Ministry of Education (1994) curriculum framework. The New Zealand Ministry of Education (2007) updated the curriculum and positioned reading comprehension with a new set of categories (i.e., process and strategies, purposes and audiences, content and ideas, language features, structure). Expert content analysis (e.g., Does item 33 of A map onto Q?) is required to determine if the definition and operationalization of A is the same as Q. My suspicion is that this matters more to psychometricians than curriculum developers. Furthermore, at least in New Zealand, the New Zealand Ministry of Education has proceeded with a curriculum review during the government mandated lockdown. Consequently, the pressure to change badges on test reports without validating the mapping of old items and categories to new will impose significant pressures on reporting and equating of reports over time.

Covid has reminded us that reporting for school or policy administrators is very different to reporting for classroom teachers. Teachers need to know 'now' who needs to be taught what next so that they can adjust lesson plans, student groupings, or activities. Putting test reports in the hands of administrators is unlikely to lead to real classroom change. Indeed, a survey study in New Zealand showed that as administrators took over the use of a test system for school evaluation purposes, most teachers saw the tests as irrelevant for classroom use (Brown and Harris, 2009). Designing tests that prioritize teacher needs, while not ignoring those of administrators, is the ambition of educational testing (i.e., helping teachers teach better; Popham, 2000).

SS: There have been various types of impact. Test publishers now must include more caveats in score reports. For example, they may need to include caveats about state averages being based on smaller percentage of students since some years of pandemic led to more limited participation in testing. There are now more "holes" in reports such as score histories and student growth score reports because of missing test score data in spring 2020 and, in many cases, spring 2021. A big problem now is the determination and reporting of the loss in learning due to COVID and the determination of ways to estimate growth even with the loss. Fortunately, there is already quite a bit of research on these issues including Gajderowicz et al. (2022), Maldonado and De Witte (2020), and Toker (2022).

Given that many tests, since the start of the pandemic, now provide remote testing options, at least two questions related to score reporting are: (1) Whether and how should one consider the testing mode while score reporting? (2) Should there be separate reporting scales for test-center examinees versus remote examinees?

Test publishers may also have to take account of the mode when choosing a norm group. For example, consider a score report that shows the scaled score along with an average performance range

[2] https://en.wikipedia.org/wiki/2020_United_Kingdom_school_exam_grading_controversy

(APR). If a test-taker took a test remotely, how should the APRs be computed? Should we compute the APRs based on only remote test-takers or should we combine both remote and test center test-taker, or report two sets of APRs? For many tests, remote test-takers perform better than test-center test-takers. So, if we compute APRs using only remote test-takers, then a particular remote test-taker will appear worse (relative to the APRs) than if we computed the APRs using both remote and test-center test-takers.

> Q2. How will the nature of research in the area of score reporting change due to the availability of data from digital learning and assessment systems?

GTLB: Work with schools makes it clear that there is no one data management system being used to handle student data including achievement or performance data. That means designing test reports that contextualize performance with useful information from background data is extremely messy and difficult. However, it may be possible to identify meaningful information, such as opportunity to learn data (e.g., was the student present when *xyz* was taught? And if not, what reason was there for the absence?), that sheds light on performance and supports appropriate responses by the teacher. Unfortunately, the cost of developing robust systems is such that there is resistance to open access between proprietorial systems that will delay the reporting of information that might help teachers understand why some students are not doing as well as expected.

A major challenge in score reporting is the churn in teacher recruitment, retirement, resignation, or transfer. This impacts the overall level of teacher assessment literacy (Xu and Brown, 2016). Although a test report system can be well designed to communicate effectively with teachers, that does not guarantee that post-deployment of the system, all teachers will be competent to understand the reports. There is a constant need to deploy support at the moment when teachers need to make decisions based on test reports. It is possible to forget what had been taught in initial teacher education (if anything was) and infrequent use of the reports creates challenges. Hence, test reporting systems need to include a variety of instructional and support resources that help teachers interpret reports correctly. Assuming that the test manual is enough simply is not warranted. The expression *RTFM* exists for a reason; users rarely consult the manual (Blackler et al., 2016). Multiple communication channels are needed to ensure teachers interpret reports as they ought to (Brown, 2019) and this must be maintained throughout the life of the test reporting system, a matter of potential economic impact as well.

SS: The nature of research will change (or maybe expand is a more appropriate word than change) in several ways. For example, it is possible to obtain additional data (e.g., timing data, key stroke data, eye tracking data, etc.) from digital assessment systems that are not available when tests are given on paper. Thus, more research and operational analyses (e.g., analysis on motivation, new test security analysis, new types of speededness analysis, etc.) can be done now that could not be done previously. Many tests now routinely conduct timing analysis and flag examinees if something appears suspicious, for example test taker(s) completing a 120-item test in 5 min (if they scored high, then they may have cheated, while if they scored low, they may have lacked motivation). Exciting research is being performed by people like Hongwen Guo and Kadriye Ercikan at ETS (Guo and Ercikan, 2021) and by Steven Wise at NWEA (Wise, 2017, 2021).

If a digital learning and assessment system can accurately measure learning progression, I suppose we have to explore ways to report the progression in an easily comprehensible manner. Some exciting work on this area has been done both at ETS by people like Aurora Graf and Peter van Rijn and outside ETS by people like Derek Briggs, but there is scope for a lot of further work. Collecting validity evidence for the utility of score reports is more difficult since we cannot just have focus groups reacting to static reports but have to have potential users navigate online, interactive reporting systems that offer different buttons/menus/choices for users to select in order to see how reports work.

PK: Online and digital learning environments, instructional technologies, and game-based learning environments have seen a rise in recent years (e.g., Heffernan and Heffernan, 2014; Feng et al., 2018; Sinatra et al., 2020; Rahimi and Shute, 2021). Students in several districts across the United States now have district-provided laptop and tablet devices which gives them access to various types of digital and online learning platforms. Their interactions in such online learning and testing environments result in much underlying background and log data such as: number of times a student accesses various features within the learning or assessment environment, where and when the student clicks, how the student navigates within the digital environment, the amount of time a student spends on the assigned task or question, the number of attempts a student makes to answer a question correctly, the number of hints and scaffolds that the student uses, to name just a few examples. Such data could provide a richer context and additional insights on a student's current state of understanding and could provide some opportunities to support more effective and personalized learning experiences for each student.

There is a great opportunity to provide feedback that is instructionally useful with the large amounts of data that can now be available in these digital contexts. There is already some interesting work in this area, and several Learning Analytic Dashboards (LADs) are being designed (Molenaar and Knoop-van Campen, 2018; Michaeli et al., 2020; Sahin and Ifenthaler, 2021) with the intent of providing real-time feedback on instructionally informative data such as students' time on task, progress toward goals, their overall level of conceptual understanding, and their strengths and needs relative to ongoing formative goals. This information can be scaffolded and presented to teachers in a real-time actionable dashboard (examples in Kannan et al., 2019; Kannan and Zapata-Rivera, 2022) with scaffolds and visualizations that can help teachers in tailoring their instruction to fill specific gaps in students' conceptual understandings.

However, with the overwhelming amount of data available, teachers are often left "data rich and information poor" (DRIP). The concept of DRIP was first extended to education about 10 or so years ago (Charman, 2009) when educators were beginning to get bombarded with large volumes of data from large-scale assessment and reporting systems. Such large amounts of data result in an unwanted increase in teachers' cognitive load when they are already strapped for time. It could also lead to several "curiosity-driven explorations" (Khosravi et al., 2021, pp. 3; Wise and Jung, 2019) of irrelevant questions, which again poses unwanted and unrequited demands on their limited time.

Therefore, it is very important that the data provided to educators is not overwhelming, and that the score reports and dashboards be designed in such a way that the information is scaffolded in an interpretable and usable way to suit the needs of the users (Kannan and Zapata-Rivera, 2022). There are a number of ways in which big data can be scaffolded. One example is to use a question-based interaction format (VanWinkle et al., 2011), where questions that reflect the needs of the intended user group are unraveled through stakeholder-specific needs assessments. The dashboards are then designed such that users can pose specific questions based on their needs and pre-canned visualizations and actionable data chunks that support instructional decision-making are populated to support their immediate use (Zapata-Rivera, 2021; Kannan et al., 2022). LADs are increasing across the educational landscape (Papamitsiou and Economides, 2014; Sahin and Ifenthaler, 2021), and research in this area should continue to focus on how these dashboards can provide timely, interpretable, usable, actionable, and useful information to educators so that they do not pose a demand on their already limited time.

AZ: This is an exciting time to be in testing. It really is. Digital learning and digital assessment systems mean data, and data means, theoretically, more and different things to say about test-taker performance (DiCerbo, 2020). But we are not quite *there* yet, especially with respect to reporting. In terms of reports, we are still in the potential stage. In part, that is because of our training and nature as professionals: many of us in reporting grew up as psychometricians who by nature are very careful about what inferences are right and proper (in terms of reliability and validity) and which are not (poetically described by Sireci (2021) as "psychometric paralysis"). The kinds of data and volumes of data we can gather are still not well understood in terms of our established validity paradigms, and this has implications for communication of those data elements.

But, with uncertainty, I think there is opportunity. We can do research in terms of big data. What does it all mean? We can do research on how to communicate these new kinds and quantities of data - how do we display different data so that differentiated user groups can be supported in various informational and actionable goals (Hegarty, 2011; Zenisky, 2015)? And not just how do we display information but also, how might we structure the data to engage users relative to those goals? That's where we get into stakeholder research on the use of reports and dashboards and connect those research activities to what users actually do with the data (e.g., Wainer et al., 1999; Rick et al., 2016; Corrin, 2018).

To that end, we still have much to learn about the kinds of questions stakeholders might have and actions they might take in light of different kinds of data that is emerging. How do we package reports and results so that they can do what they need to do, in terms of anticipating those needs? That is one direction where reporting research could be heading. Some user questions will be informational in nature, some will be actionable in nature, and thus the task in front of us is to learn enough to build the systems that let people get what they need to get, to do what they need to do.

> Q3. What aspects of your work can inform current work on designing and evaluating interactive reports and dashboards?

SS: Interactive reports and dashboards aim to produce a wealth of information about test-taker's engagement, experiences, and performance on tasks (Kannan and Zapata-Rivera, 2022). Users of interactive reports and dashboards most likely would want to see the results on various aspects of learning and various types of interaction between a test-taker and the system, on the difference in performance, of the same group over multiple time periods, over multiple groups for the same time period, over multiple groups over several time periods, and often at the subscale level. My research on score reporting has focused mainly on subscores (Sinharay et al., 2011) and more generally on evaluating psychometric quality of the reported information (Sinharay and Johnson, 2019; Sinharay, 2021). So, I suppose my work may be relevant in the determination of:

- the information that is appropriate to include in interactive reports and dashboards,
- the appropriate interpretations of the information, and
- the appropriate uses of the information.

And I anticipate that my research may be helpful in answering questions like:

- Is it justified to report all the information that is intended to be reported or is demanded by clients?
- Are the various scores reliable enough to be reported or interpreted for their intended purposes?
- For situations when various scores for the individual examinees are not reliable enough, are the differences between average school-level scores reliable enough to be reported?

PK: In my current role at WestEd, I work mostly with state departments of education and regional educational laboratories in providing consultation, technical assistance, and other psychometric support particularly in areas related to score reporting. Assessment programs that we have recently been supporting for several states are in the early learning space (basically preschool through kindergarten assessments that assess children in broad domains such as social–emotional skills, independence, and motor coordination, and foundational knowledge and skills that prepare them for kindergarten). These assessments are often administered to fulfill the reporting requirements for the Office of Special Education Programs (OSEP), which requires that children with disabilities who enter a publicly funded preschool education program are assessed upon entry and exit to the program to demonstrate growth in the assessed domains. These assessments can also be used to monitor progress over time for all children enrolled in publicly funded preschool programs, but the OSEP reporting use-case is the most common as it is a federal reporting requirement in those states.

There are several challenges in assessing and reporting for this age group and this population. Educators in preschool contexts are often less familiar with large-scale assessments (Pretti-Frontczak et al., 2002; Ertle et al., 2016; Schachter et al., 2019) and reports from these large-scale assessments. Administering standardized assessments (which are often observation-based) in this context comes with its own set of challenges (Finello, 2011). Parents are often unfamiliar with the need for and context of assessment in this space. We have encountered several challenges in developing appropriate score reports and dashboards that are interpretable and useful for educators and parents. I can provide a few examples here to illustrate.

- Scale scores often do not mean anything to these stakeholders, particularly parents of preschoolers, and they are often baffled by such numbers on reports. So, we have been working with our state department clients to identify ways in which we could report results from multiple interim assessments without indicating a scale score, but in such a way that parents can still see progress or movement made by their children.

- Teachers and parents want benchmarks and normative comparisons to compare their child's performance. As you can imagine, at this age group, stakeholders, particularly parents, are used to percentiles. Parents often take their child to the pediatrician's office, and are told that their child is 25th percentile in height, 50th percentile in weight, and so on. But, though these early learning assessments are based on underlying learning progressions, these assessments have not been normed and do not have normative samples to show such comparisons. Therefore, we have been working with our clients to identify other criterion-based benchmarks (e.g., average level for 3- and 5-year-olds, or average level that indicates kindergarten readiness) to provide benchmark comparisons for stakeholders.
- During a needs assessment study for a couple of the client states, we found that teachers and parents often want feedback at a nuanced skill-level, but reporting at this individual skill (item) level is not feasible when these measurements are often based on a single time-point and a single observation. We have been working with our clients to provide sufficiently detailed feedback that may be useful in an interim assessment context, while at the same time not providing item-level details that can lead to over-interpretation/misinterpretation of the results.
- Teachers/speech language pathologists and others who use these data often find the reporting dashboards confusing and overwhelming to navigate and to use in their practice. Our evaluations with teachers have revealed additional needs for professional development (PD) to understand and use the reporting dashboards. We have been working with our partners and clients to identify and create appropriate PD videos that can be accessed by teachers at any time.

Keeping the stakeholder (user) at the center of the design and evaluation process (Kannan, 2023), we are also working with teachers and parents in these states by implementing the iterative multi-step approach (Zapata-Rivera et al., 2012; Hambleton and Zenisky, 2013) and repeatedly evaluating iterations of the Individual Student Reports and dashboards with the intended stakeholders to evaluate the extent to which they are able to accurately interpret and appropriately use the results (Kannan, 2023). Parents are a particularly heterogeneous stakeholder group (Kannan et al., 2018, 2021), and it is critically important to ensure that the reports designed for parents provide them with interpretable information that they can appropriately use. Therefore, the extent to which our evaluations indicate gaps in interpretation and/or use by teachers and parents, we have been making additional revisions to scaffold and elucidate the information being provided to specifically cater to stakeholder needs, and cycle back in for additional rounds of evaluations.

AZ: At present, I'm working on several different projects in the adult education space that involve both formative and summative assessments, and in each project, I am closely engaged with teachers. I keep hearing the word "actionable" being used to describe assessments and assessment results, and that to me is a difficult-to-pin-down term if we consider that reporting traditionally is a top-down activity. To that end, again, there is a real need to view stakeholder groups where they are in terms of assessment literacy and also feedback literacy and recognize the presence of variability even within groups. Users *must* be centered in this process. In the idea of actionable there is a potential for interactivity in reporting, and that to me suggests building tools that do not necessarily answer specific questions but rather flexibly respond to the kinds of questions stakeholders might pose (Zenisky and Hambleton, 2013; Zenisky, 2016).

In terms of designing and evaluating interactive reports and dashboards, it's perhaps better from an efficiency perspective to have the programmers jump right to the code and build, but that likely means the specifications for the build are based on something other than users. I was in a meeting recently where a very nice project manager wanted a very specific list of changes to a report to hand to a developer to check a box on a deliverable, and they were perhaps less than thrilled when I declined to provide such a list. In that case, I myself am nowhere near the target user of the assessment. I can critique reports, and I can advise on what the literature speaks to as good practice generally, but the stakeholders are the ones we need to listen to (Zenisky and Hambleton, 2015; Corrin, 2018). That's where my current work in reporting is now. Defining "actionable" and using the words of stakeholders to guide reporting, rather than "you'll take what I give you and like it."

GTLB: My own research (Brown, 2004, 2008) into how teachers conceive of the purposes and nature of assessments, including tests, shows that their pre-existing conceptions matter to how they use tests. When the priority is on using assessment to improve learning, then informal formats are prioritized, especially if they reveal insights into students' deeper or higher-order learning (Brown, 2009). However, survey research with students showed that students who believe teachers use formal testing to improve teaching performed better (Brown et al., 2009). Further, students who accept that testing will legitimately evaluate their accomplishments tend to do better (Brown, 2011). Not surprising, how teachers understand assessment matters to their behaviors. Thus, how assessment is designed in any jurisdiction matters to how teachers and students will understand and respond to assessment (Brown and Harris, 2009). Environments are not equal, so there are few universalities in how teachers conceive of assessment (Brown et al., 2019), meaning that test reports cannot be universal either.

The pre-existing belief structures of teachers and administrators have been derived from their extended experience of assessments in formal and informal environments. This means that test reports must be designed in light of those factors (Brown and Hattie, 2012). Test reports must go beyond total score and rank order to provide usable instructional insights. Notwithstanding concerns about the validity and reliability of sub-score reports (Sinharay et al., 2018), sub-score reports from well-designed tests provide a more robust basis for decision making than teachers' own intuitions. This means that teachers need guidance from test reports to reduce the temptation to believe students' skill is less than what a well-designed test shows they can do. This was demonstrated in New Zealand where teachers judged almost all children to be worse than what their test performance showed (Meissel et al., 2017). For formative purposes, it is more likely that sub-scores will be educational.

Moreover, teachers have theories that explain why some students learn and others do not (e.g., *students do not learn because of poverty*).

Data-driven decision-making professional development strategies (Lai and Schildkamp, 2016) seem to help teachers re-examine their *a priori* beliefs about poor performance. Test reports or data visualizations that help teachers see that their favored explanation is invalid (e.g., scores by students with or without indicators of poverty) are needed so that teachers and leaders can grapple with the responsibilities and authority they have to ensure learning occurs. Without reports that provide information that helps teachers and leaders do their jobs and which simultaneously correct wrong thinking about learning, tests will not do much to improve equitable outcomes.

> Q4. What new challenges and opportunities you expect will be present in the field of score reporting?

AZ: Reporting, when we think about it, goes to the heart of the issue of why we do what we do – it facilitates the 'why' of testing. Right now, however, public appetite for summative tests that do not say much at a granular level is waning quickly. Quite honestly, sometimes it does feel like there are nuances to testing and test scores that only testing professionals care deeply about. So, for those of us who work in reporting, we may at times find ourselves in a netherworld of sorts, where we understand all too well the technicalities of data and psychometrics, while at the same time having one foot firmly in the world of users and real-world use of tests. That is *a* challenge, to figure out how to talk about test results considering the reality of the data and all the uncertainty that comes with it. But this is also an opportunity.

Relatedly, another challenge has been, and will be in the coming years, how to make tests relevant. When we say that a test cannot be used for this, or is not valid for that, despite taking hours of time to administer, and occupying a huge amount of mental space for educators, those tests are being relegated to noise. I believe we do need to push harder on being realistic about what specific tests can and cannot do, and advance reporting methods appropriate to different tests.

Lastly, I think a new challenge will be how to navigate the push for artificial intelligence as a way to solve everything and ensure that people are still involved in the process of reporting. There are certainly ways that technology can be involved in our work and that can help us with some of the tricky bits, but there still remains no substitute for the engagement with stakeholders and iterations that spring from that that are informed by real users and use cases (Kannan et al., 2018; Slater et al., 2018).

GTLB: As a research scientist, I am excited by the many intriguing possibilities generated by new technologies to examine what students are doing while they answer online tests. Technologies such as eye-tracking, process logs, galvanic skin responses, event related potentials, among others have the potential to reveal what the mind is really doing. Vast amounts of data can be generated by these technologies some of which will no doubt correlate with tested performance. However, there is a strong possibility that the associations of eye movements, skin responses, mouse usage, or brain electrical activity are not related with any meaningful principles that could inform instruction. A similar debate has been held around *f*MRI studies (Vul et al., 2009).

It was only when Greiff et al. (2015) developed a conceptual framework of how a problem should be solved under the scientific method that they could make sense of how students responded to an online test of reasoning. A recent study of process data on a computer-based test made sense only when time spent on an item and actions within the item were interpreted as evidence of persistent effort; analysis showed that effortful persistence contributed to overall better performance even when a specific item was answered incorrectly (Lundgren and Eklöf, 2020). These studies, among others, show that the data by themselves do not make sense of themselves (Pearl and Mackenzie, 2018). Zumbo et al. (2023) have identified the key issue with the promise of sensor data; it lacks a coherent psychological theory to explain how the movement of eyes, mice, electricity, or blood relates to instruction and learning. As a discipline, we cannot explain why these physiological or behavioral data mean anything for understanding teaching and learning. Thus, much needs to be done to not just display sensor data but communicate how that information can be usable by teachers and administrators. Does it matter? For now, we do not know.

SS: As I mentioned earlier, one challenge, given the availability of an ocean of data, is to determine exactly how much information from data on timing, eye-tracking, learning, growth, and so on is justified to be reported from the viewpoints of accuracy, reliability, or validity. This determination will require both psychometric analysis as well as focus groups, discussions with parents and teachers, and so on.

One challenge is that we have to improve data literacy as consumers of test scores get overwhelmed with more and more data/test score information from summative and interim assessments. An opportunity is that we could report a substantial amount of new information that would provide a context for interpreting a score. For K-12 assessments, for example, we could flag a score where the student appeared unmotivated (i.e., lots of omits, fast guessing, low engagement characteristics, etc.). A teacher (or parent) should know that a score under these conditions may not represent the student's best work. For writing samples, one may report flags indicating that a student used poor strategies in production (e.g., little revision, no signs of outlining). I admit, though, that these options could lead to challenges/lawsuits by the test takers against the testing company, similar to what testing companies fear tests-takers might do when their scores are put on hold or canceled due to possible unfair practices, such as cheating.

I mentioned research on reporting loss of learning due to the pandemic. An opportunity is to go one step further and conduct research on a more general topic—how to report and what to report when a major disruption or unexpected problem occurs, where an unexpected problem could be a disease outbreak, or could be a natural disaster, a huge computer problem, or something different.

PK: I agree with all of the points made by my fellow panelists here. From my perspective, I think that all of my previous responses are somewhat related to the challenges and opportunities that we are faced with, in the field of score reporting at this time.

First, within the current political climate and the overall push-back on assessments, score reports can be used as a vehicle for pushing the needle toward a conversation around educational equity by changing the narrative from 'achievement gaps' to 'opportunity gaps' and beginning to address the systemic issues in opportunities and access that persist across various racial and class lines. Score reports can and should be used as an effective tool in unraveling these systemic issues by not only using OTL variables for disaggregating data, but also by conducting intersectional analyses to identify bias and using effective visualizations to clearly report these results to various stakeholder groups so that it results in much needed action. There are likely to be measurement challenges associated with these types of analyses, such as the over-interpretation of correlational results in associating OTL with educational outcomes. Still, this shift in focus could be the necessary first step in helping policymakers in identifying the opportunity gaps and enabling them to start addressing systemic issues through appropriate legislations.

Second, as I pointed out earlier, the availability of large amounts of data, such as time spent on task, number of attempts made, number of hints or scaffolds used, can have its own associated benefits and challenges. While such data can be promising in providing instructionally useful feedback to teachers and personalized learning experiences for students, it is important to ensure that such data is appropriately scaffolded for the users so that teachers and students are able to understand the results presented and take appropriate actions from these results. Using techniques such as the 'question-based interaction approach' could help scaffold this data for end-users making it easy to take appropriate actions.

Finally, the context-specificity and stakeholder-specificity of score reports and dashboards (Kannan, 2023) is further exemplified by the early learning contexts we currently work with at WestEd. This highlights the critical importance of using an audience-specific approach and implementing an iterative multistep framework (Zapata-Rivera et al., 2012; Hambleton and Zenisky, 2013) in the design and evaluation of dashboards to create reports that are interpretable and useful to the end users.

Conclusion

We have discussed several challenges and opportunities in the field of score reporting that result from changes in society and education. Some of these challenges and opportunities have to do with alignment between assessment type and appropriate use. This includes the types of decisions the assessment supports and the assessment information that it produces. Our experts elaborated on the challenges that COVID has imposed regarding the nature of research and practice on score reporting. These include: the need to shift away from reporting scores to providing insights that have clear relevance for instruction (e.g., what teachers need to know now to support the next instructional activities), co-designing reports with the intended audience considering equity and context factors, recognizing the different purposes of assessments and their intended use, and the implications for reporting assessment results due to test mode (e.g., remote vs. in-person testing).

Regarding the amount of data available using digital learning assessment and learning systems, the experts mentioned the opportunities that additional multimodal data will afford (e.g., assessing motivation and detecting cheating) and warned us about the need for developing appropriate methods that help us analyze rich process data to support decision making. Also, the authors elaborated on the importance of designing reporting systems (or dashboards) that take into account the needs of the audience and offer support for appropriate interpretation (Kannan and Zapata-Rivera, 2022).

The experts' current research can inform new developments in the field of score reporting in various ways including identifying information to include in interactive reports and dashboards, co-designing and evaluating new reporting systems with the audience to ensure the reports provide relevant insights, exploring the psychometric properties of assessment results, developing materials to support teacher interpretation of assessment results.

Finally, the experts consider that challenges mentioned can be opportunities for the creation of assessments that provide relevant insights for different audiences. For example, new psychometric methods that can deal with disruptive situations will be developed. It is expected that conversations around educational equity will impact the way reports are designed and evaluated.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

GB: conceptualization, writing—original draft, writing—review and editing, and funding acquisition. PK, SS, and AZ: writing—original draft. DZ-R: methodology, supervision, writing—original draft, and writing—review and editing. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Alessio, H. M., Malay, N. J., Maurer, K., Bailer, A. J., and Rubin, B. (2017). Examining the effect of proctoring on online test scores. *Online Learning J.* 21. doi: 10.24059/olj.v21i1.885

Bailey, D. H., Duncan, G. J., Murnane, R. J., and Au Yeung, N. (2021). Achievement gaps in the wake of COVID-19. *Educ. Res.* 50, 266–275. doi: 10.3102/0013189X211011237

Blackler, A. L., Gomez, R., Popovic, V., and Thompson, M. H. (2016). Life is too short to RTFM: how users relate to documentation and excess features in consumer products. *Interact. Comput.* 28, 27–46. doi: 10.1093/iwc/iwu023

Brown, G. T. L. (2004). Teachers' conceptions of assessment: implications for policy and professional development. *Assess. Educ. Princip. Policy Pract.* 11, 301–318. doi: 10.1080/0969594042000304609

Brown, G. T. L. (2008). *Conceptions of assessment: Understanding what assessment means to teachers and students*. Hauppauge: Nova Science Publishers.

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

Brown, G. T. L. (2011). Self-regulation of assessment beliefs and attitudes: a review of the Students' conceptions of assessment inventory. *Educ. Psychol.* 31, 731–748. doi: 10.1080/01443410.2011.599836

Brown, G. T. L. (2019). Technologies and infrastructure: costs and obstacles in developing large-scale computer–based testing. *Educ. Inq.* 10, 4–20. doi: 10.1080/20004508.2018.1529528

Brown, G. T. L., Gebril, A., and Michaelides, M. P. (2019). Teachers' conceptions of assessment: a global phenomenon or a global localism. *Front. Educ.* 4:e0016. doi: 10.3389/feduc.2019.00016

Brown, G. T. L., and Harris, L. R. (2009). Unintended consequences of using tests to improve learning: how improvement-oriented resources heighten conceptions of assessment as school accountability. *J. MultiDisc. Eval.* 6, 68–91. doi: 10.56645/jmde.v6i12.236

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

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

Brown, G. T. L., Peterson, E. R., and Irving, S. E. (2009). "Beliefs that make a difference: adaptive and maladaptive self-regulation in students' conceptions of assessment" in *Student perspectives on assessment: what students can tell us about assessment for learning*. eds. D. M. McInerney, G. T. L. Brown and G. A. D. Liem (Charlotte: Information Age Publishing)

Charman, P. (2009). Data rich, information poor: creative and innovative approaches to results analysis to support teaching and learning. Paper presented at the 35th Annual Conference of the International Association for Educational Assessment, Brisbane, Australia.

Corrin, L. (2018). Evaluating students' interpretation of feedback in interactive dashboards. In D. Zapata-Rivera (Ed.), *Score reporting: research and applications* (pp. 145–159). London: Routledge.

Dawson, P. (2021). *Defending assessment security in a digital world: preventing e-cheating and supporting academic integrity in higher education*. London: Routledge).

DiCerbo, K. (2020). Assessment for learning with diverse learners in a digital world. *Educ. Meas. Issues Pract.* 39, 90–93. doi: 10.1111/emip.12374

Dorn, E., Hancock, B., Sarakatsannis, J., and Virulea, E. (2020). COVID-19 and student learning in the United States: the hurt could last a lifetime. McKinsey & Company. Available at: https://www.childrensinstitute.net/sites/default/files/documents/COVID-19-and-student-learning-in-the-United-States_FINAL.pdf (Accessed April 19, 2023).

Ertle, B., Rosenfeld, D., Presser, A. L., and Goldstein, M. (2016). Preparing preschool teachers to use and benefit from formative assessment: the birthday party assessment professional development system. *Math. Educ.* 48, 977–989. doi: 10.1007/s11858-016-0785-9

Feng, M., Krumm, A., and Grover, S. (2018). "Applying learning analytics to support instruction" in *Score reporting research and applications*. ed. D. Zapata-Rivera (New York, NY: Routledge), 145–159.

Finello, K. M. (2011). Collaboration in the assessment and diagnosis of preschoolers: challenges and opportunities. *Psychol. Sch.* 48, 442–453. doi: 10.1002/pits.20566

Gajderowicz, T. J., Jakubowski, M. J., Patrinos, H. A., and Wrona, S. M. (2022). Capturing the educational and economic impacts of school closures in Poland. 10253Policy Research Working Paper Series, The World Bank.

Goodman, D., and Hambleton, R. K. (2004). Student test score reports and interpretive guides: review of current practices and suggestions for future research. *Appl. Meas. Educ.* 17, 145–220. doi: 10.1207/s15324818ame1702_3

Goudeau, S., Sanrey, C., Stanczak, A., Manstead, A., and Darnon, C. (2021). Why lockdown and distance learning during the COVID-19 pandemic are likely to increase the social class achievement gap. *Nat. Hum. Behav.* 5, 1273–1281. doi: 10.1038/s41562-021-01212-7

Greiff, S., Wüstenberg, S., and Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Comput. Educ.* 91, 92–105. doi: 10.1016/j.compedu.2015.10.018

Guo, H., and Ercikan, K. (2021). Comparing test-taking behaviors of English language learners (ELLs) to non-ELL students: use of response time in measurement comparability research (research report no. RR-21-25). ETS. doi: 10.1002/ets2.12340,

Hambleton, R., and Zenisky, A. (2013). *Reporting test scores in more meaningful ways: a research-based approach to score report design. APA handbook of testing and assessment in psychology*. Washington, DC: APA.

Hattie, J. A., and Brown, G. T. L. (2008). Technology for school-based assessment and assessment for learning: development principles from New Zealand. *J. Educ. Technol. Syst.* 36, 189–201. doi: 10.2190/ET.36.2.g

Heffernan, N., and Heffernan, C. (2014). The ASSISTments ecosystem: building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *Int. J. Artif. Intell. Educ.* 24, 470–497. doi: 10.1007/s40593-014-0024-x

Hegarty, M. (2011). The cognitive science of visual-spatial displays; implications for design. *Top. Cogn. Sci.* 3, 446–474. doi: 10.1111/j.1756-8765.2011.01150.x

Hegarty, M. (2018). "Advances in cognitive science and information visualization" in *Score reporting research and applications*. ed. D. Zapata-Rivera (London: Routledge).

Kannan, P. (2023). "Score reporting: design and evaluation methods informed by research" in *International encyclopedia of education*. eds. R. J. Tierney, F. Rizvi and K. Ercikan. *4th* ed (Amsterdam: Elsevier)

Kannan, P., Beigman-Klebanov, B., Shao, S., and Long, R. (2019). Evaluating teachers' needs for on-going feedback from a technology-based book reading intervention. Paper presented at the 2019 Annual Meeting of the National Council for Measurement in Education, Toronto, ON.

Kannan, P., Deane, P., and Phelps, G. (2022). Validating writing traits for classroom assessment purposes. Paper presented at the 2022 annual meeting of the American Educational Research Association.

Kannan, P., and Zapata-Rivera, D. (2022). Facilitating the use of data from multiple sources for formative learning in the context of digital assessments: informing the design and development of learning analytic dashboards. *Front. Educ.* 7:913594. doi: 10.3389/feduc.2022.913594

Kannan, P., Zapata-Rivera, D., and Bryant, A. D. (2021). Evaluating parent comprehension of measurement error information presented in score reports. *Practical Assessment, Evaluation, & Research.* 26. doi: 10.7275/rgwg-t355

Kannan, P., Zapata-Rivera, D., and Leibowitz, E. A. (2018). Interpretation of score reports by diverse subgroups of parents. *Educ. Assess.* 23, 173–194. doi: 10.1080/10627197.2018.1477584

Khosravi, H., Shabaninejad, S., Bakaria, A., Sadiq, S., Indulska, M., and Gasevic, D. (2021). Intelligent learning analytics dashboards: automated Drill-down recommendations to support teacher data exploration. *J. Learn. Anal.* 8, 133–154. doi: 10.18608/jla.2021.7279

Krause, K. H., Verlenden, J. V., Szucs, L. E., Swedo, E. A., Merlo, C. L., Niolon, P. H., et al. (2022). Disruptions to school and home life among high school students during the COVID-19 pandemic - adolescent behaviors and experiences survey, United States, January-June 2021. *Morbid. Mortal. Week. Rep. 1* 71, 28–34. doi: 10.15585/mmwr.su7103a5

Ladson-Billings, G. (2006). From the achievement gap to the education debt: understanding achievement in U. S. Schools. 2006 presidential address. *Educ. Res.* 35, 3–12. doi: 10.3102/0013189X035007003

Lai, M. K., and Schildkamp, K. (2016). "In-service teacher professional learning: use of assessment in data-based decision-making" in *Handbook of human and social conditions in assessment*. eds. G. T. L. Brown and L. R. Harris (London: Routledge)

Lundgren, E., and Eklöf, H. (2020). Within-item response processes as indicators of test-taking effort and motivation. *Educ. Res. Eval.* 26, 275–301. doi: 10.1080/13803611.2021.1963940

Mahnken, K. (2022). 'Nation's report card': two decades of growth wiped out by two years of pandemic. The 74 million. Available at: https://www.the74million.org/article/nations-report-card-two-decades-of-growth-wiped-out-by-two-years-of-pandemic/ (Accessed April 19, 2023).

Maldonado, J., and De Witte, K. (2020). The effect of school closures on standardised student test outcomes. Discussion paper series, no. DPS20.17, KU Leuven Department of economics, Leuven Available at: https://feb.kuleuven.be/research/economics/ces/documents/DPS/2020/dps2017.pdf (accessed on 13 October 2020).

Marion, S. (2020). Using opportunity-to-learn data to support educational equity. Center for Assessment. Available at: https://www.nciea.org/wp-content/uploads/2021/11/CFA-Marion.OTL_.Indicators.pdf (Accessed April 19, 2023).

McDonnell, L. M. (1995). Opportunity to learn as a research concept and a policy instrument. *Educ. Eval. Policy Anal.* 17, 305–322. doi: 10.3102/01623737017003305

Meissel, K., Meyer, F., Yao, E. S., and Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: exploring student characteristics that influence teacher judgments of student ability. *Teach. Teach. Educ.* 65, 48–60. doi: 10.1016/j.tate.2017.02.021

Michaeli, S., Kroparo, D., and Hershkovitz, A. (2020). Teachers' use of education dashboards and professional growth. *Int. Rev. Res. Open Distrib. Learn.* 21, 61–78. doi: 10.19173/irrodl.v21i4.4663

Modan, N. (2022). These 6 charts highlight COVID-19's impact on NAEP scores. K-12 dive. Available at: https://www.k12dive.com/news/these-6-charts-highlight-covid-19s-impact-on-naep-scores/634852/ (Accessed April 19, 2023).

Molenaar, I., and Knoop-van Campen, C. A. N. (2018). How teachers make dashboard information actionable. *IEEE Trans. Learn. Technol.* 12, 347–355. doi: 10.1109/TLT.2018.2851585

Murdock, T. B., Stephens, J. M., and Grotewiel, M. M. (2016). "Student dishonesty in the face of assessment: who, why, and what we can do about it" in *Handbook of human and social conditions in assessment*. eds. G. T. L. Brown and L. R. Harris (London: Routledge).

New Zealand Ministry of Education (1994). *English in the New Zealand curriculum.* Wellington, NZ: Learning Media.

New Zealand Ministry of Education (2007). *The New Zealand curriculum for English-medium teaching and learning in years 1–13.* Wellington, NZ: Learning Media.

O'Donnell, F., and Sireci, S. G. (2018). "Score reporting issues for licensure, certification, and admissions programs" in *Score reporting research and applications*. ed. D. Zapata-Rivera (London: Routledge).

O'Donnell, F., and Zenisky, A. L. (2020). Digital module 21: results reporting for large-scale assessments. *Educ. Meas. Issues Pract.* 39, 137–138. doi: 10.1111/emip.12408

Papamitsiou, Z., and Economides, A. (2014). Learning analytics and educational data Mining in Practice: a systematic literature review of empirical evidence. *Educ. Technol. Soc.* 17, 49–64.

Parks, S. E., Zviedrite, N., Budzyn, S. E., Panaggio, M. J., Raible, E., Papazian, M., et al. (2021). COVID-19-related school closures and learning modality changes - United States, august 1-September 17, 2021. *Morbid. Mortal. Week. Rep.* 70, 1374–1376. doi: 10.15585/mmwr.mm7039e2

Pearl, J., and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. New York: Hachette Book Group.

Popham, W. J. (2000). *Modern educational measurement: Practical guidelines for educational leaders* (*6th*). Boston: Allyn & Bacon.

Porter, A. C. (1995). The uses and misuses of opportunity-to-learn standards. *Educ. Res.* 24, 21–27. doi: 10.2307/1176117

Pretti-Frontczak, K., Kowalski, K., and Brown, R. D. (2002). Preschool teachers' use of assessments and curricula: a statewide examination. *Except. Child.* 69, 109–123. doi: 10.1177/001440290206900108

Rahimi, S., and Shute, V. (2021). "Learning analytics dashboards in educational games" in *Visualizations and dashboards for learning analytics. Advances in analytics for learning and teaching*. eds. M. Sahin and D. Ifenthaler (Cham: Springer).

Reisenwitz, T. H. (2020). Examining the necessity of proctoring online exams. *J. High. Educ. Theory Pract.* 20. doi: 10.33423/jhetp.v20i1.2782

Rick, F., Slater, S., Kannan, P., Sireci, S. G., Dickey, J., and Zenisky, A. L. (2016). Parents' perspectives on summative test score reports. Center for Educational Assessment Research Report no. 937. Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Sahin, M., and Ifenthaler, D. (2021). "Visualization and dashboards: challenges and future directions" in *Visualizations and dashboards for learning analytics. Advances in analytics for learning and teaching*. eds. M. Sahin and D. Ifenthaler (Cham: Springer)

Schachter, R. E., Strang, T. M., and Piasta, S. B. (2019). Teachers' experiences with a state-mandated kindergarten readiness assessment. *Early Years* 39, 1–17. doi: 10.1080/09575146.2017.1297777

Sinatra, A., Graesser, A. C., Hu, X., Goldberg, B., and Hampton, A. J. (2020). Design recommendations for intelligent tutoring systems: Volume 8-data visualization. US Army Combat Capabilities Development Command–Soldier Center.

Sinharay, S. (2021). Score reporting for examinees with incomplete data on large-scale educational assessments. *Educ. Meas. Issues Pract.* 40, 79–91. doi: 10.1111/emip.12396

Sinharay, S., and Johnson, M. S. (2019). Measures of agreement: reliability, classification accuracy, and classification consistency. In DavierM. von and Y. S. Lee (Eds.), *Handbook of diagnostic classification models*. Singapore: Springer Nature.

Sinharay, S., Puhan, G., and Haberman, S. J. (2011). An NCME instructional module on subscores. *Educ. Meas. Issues Pract.* 30, 29–40. doi: 10.1111/j.1745-3992.2011.00208.x

Sinharay, S., Puhan, G., Haberman, S. J., and Hambleton, R. K. (2018). "Subscores: when to communicate them, what are their alternatives, and some recommendations" in *Score reporting research and applications*. ed. D. Zapata-Rivera (London: Routledge).

Sireci, S. G. (2021). NCME presidential address 2020: valuing educational measurement. *Educ. Meas. Issues Pract.* 40, 7–16. doi: 10.1111/emip.12415

Slater, S., Livingston, S. A., and Silver, M. (2018). "Score reports for large-scale testing programs: managing the design process" in *Score reporting research and applications*. ed. D. Zapata-Rivera (London: Routledge).

Sparks, S. D. (2022). Two decades of Progress, nearly gone: National Math, Reading Scores Hit Historic Lows. Education Week. Available at: https://www.edweek.org/leadership/two-decades-of-progress-nearly-gone-national-math-reading-scores-hit-historic-lows/2022/10 (Accessed April 19, 2023).

Tannenbaum, R. J. (2018). "Validity aspects of score reporting" in *Score reporting research and applications*. ed. D. Zapata-Rivera (London: Routledge).

Toker, T. (2022). Detecting possible learning losses due to COVID-19 pandemic: an application of curriculum-based assessment. *Int. J. Contemp. Educ. Res.* 9, 78–86. doi: 10.33200/ijcer.985992

VanWinkle, W., Vezzu, M., and Zapata-Rivera, D. (2011). *Question-based reports for policymakers (research memorandum no. RM-11–16)*. Princeton, NJ: Educational Testing Service.

Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspect. Psychol. Sci.* 4, 274–290. doi: 10.1111/j.1745-6924.2009.01125.x

Wainer, H., Hambleton, R. K., and Meara, K. (1999). Alternative displays for communicating NAEP results: a redesign and validity study. *J. Educ. Meas.* 36, 301–335. doi: 10.1111/j.1745-3984.1999.tb00559.x

Wise, S. L. (2017). Rapid-guessing behavior: its identification, interpretations, and implications. *Educ. Meas. Issues Pract.* 36, 52–61. doi: 10.1111/emip.12165

Wise, S. L. (2021). Six insights regarding test-taking disengagement. *Educ. Res. Eval.* 26, 328–338. doi: 10.1080/13803611.2021.1963942

Wise, A. F., and Jung, Y. (2019). Teaching with analytics: towards a situated model of instructional decision-making. *J. Learn. Anal.* 6, 53–69. doi: 10.18608/jla.2019.62.4

Wuthisatian, R. (2020). Student exam performance in different proctored environments: evidence from an online economics course. *Int. Rev. Econ. Educ.* 35:100196. doi: 10.1016/j.iree.2020.100196

Wyse, A. E., Stickney, E. M., Butz, D., Beckler, A., and Close, C. N. (2020). The potential impact of COVID-19 on student learning and how schools can respond. *Educ. Meas. Issues Pract.* 39, 60–64. doi: 10.1111/emip.12357

Xu, Y., and Brown, G. T. L. (2016). Teacher assessment literacy in practice: a reconceptualization. *Teach. Teach. Educ.* 58, 149–162. doi: 10.1016/j.tate.2016.05.010

Zapata-Rivera, D. (2021). Open student modeling research and its connections to educational assessment. *Int. J. Artif. Intell. Educ.* 31, 380–396. doi: 10.1007/s40593-020-00206-2

Zapata-Rivera, D., Kannan, P., and Zwick, R. (2018). "Communicating measurement error information to teachers and parents" in *Score reporting research and applications*. ed. D. Zapata-Rivera (London: Routledge).

Zapata-Rivera, D., VanWinkle, W., and Zwick, R (2012). Applying score design principles in the design of score reports for CBAL™ teachers. Research memorandum no. RM-12-20. Princeton, NJ: Educational Testing Service.

Zenisky, A. L. (2015). "Visual displays for reporting test data: making sense of test performance" in *Use of visual displays in research and testing: Coding, interpreting, and reporting data*. eds. M. McCrudden, G. Schraw and C. Buckendahl (Charlotte: Information Age Publishing).

Zenisky, A. L. (2016). Choose your own (data) adventure: perils and pitfalls, and lots of promise. *The NERA Researcher* 54, 11–14.

Zenisky, A. L., and Hambleton, R. K. (2013). "From "Here's the story" to "You're in charge": developing and maintaining large-scale online test and score reporting resources" in *Improving large-scale assessment in education*. eds. M. Simon, M. Rousseau and K. Ercikan (New York, NY: Routledge).

Zenisky, A. L., and Hambleton, R. K. (2015). "Test score reporting: best practices and issues" in *Handbook of test development*. eds. S. Lane, M. Raymond and T. Haladyna. *2nd* ed (New York, NY: Routledge).

Zenisky, A. L., Keller, L. A., and Park, Y. (2018). "Reporting student growth: challenges and opportunities" in *Score reporting research and applications*. ed. D. Zapata-Rivera (London: Routledge).

Zumbo, B. D., Maddox, B., and Care, N. M. (2023). Process and product in computer-based assessments. *Eur. J. Psychol. Assess.* doi: 10.1027/1015-5759/a000748