13

Reporting Scores and Other Results

April L. Zenisky
University of Massachusetts Amherst

Francis O'Donnell
NBME

Ronald K. Hambleton
University of Massachusetts Amherst

Zenisky, A. L., O'Donnell, F., & Hambleton, R. K. (2025). Reporting scores and other results. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 895–958). Oxford University Press.
DOI: 10.1093/oso/9780197654965.003.0013

When most people who are not test developers think about tests and measurements, it is not models or theories that typically come to mind. It is test scores and how they have been presented and explained that is memorable and often consequential. Basic questions often asked, such as "Did you pass?," "How'd you do?," and "What does your score mean?," speak to the impact that tests and test results have for stakeholders and where these users' primary interests reside. Tests, by and large, fulfill a specific purpose for specific users, and there are actions and choices to be made on the basis of the scores, such as advancement down a career path, needing to take a review course or sign up for tutoring, or applying—or not—for admission to certain institutions. For these reasons, it is critical that test takers and other score users understand what the results mean relative to whatever actions may be available given the validated interpretation(s) and use(s) of the results.

Reporting test results is thus perhaps the single most important point of interaction between the testing agency and users of test scores, and, as the old saying goes, "You never get a second chance to make a first impression." A well-designed and carefully crafted report of test results should support stakeholders by first communicating information about performance and, second, providing evidence to support whatever action(s) follow from that report of performance. In contrast, a poorly designed or confusing report does not lend itself to accomplishing an informational or actionable purpose and, in discouraging understanding of results, also can foster skepticism and/or distrust in a test.

GOALS FOR THE CHAPTER AND ADVANCE ORGANIZER

In this chapter, the intent is to provide readers with a broad theoretical context for understanding reporting and reporting choices and to draw on the growing body of research on reporting test results that is relevant to practical problems. To be clear, the approach here is less on prescriptive findings from prior studies and more on offering general trends where appropriate, tempered by the guidance for report development teams to engage with users whenever possible, to the extent possible.

The chapter begins with an overview of frames and perspectives for reporting, acknowledging that reports come in various sizes, shapes, and forms, and there are several conceptual approaches to reporting that are reflective of specific stakeholders' interests and needs. Reporting efforts and report design approaches must be explicitly and directly tied to validity, meaning they must—first and foremost—support the intended interpretations and uses of test results. From there, the focus will turn to defining reporting in terms of several key dimensions, including the unit of analysis for reporting results, assessment contexts, report contents, mechanisms for reporting, and the context of reporting efforts within broader reporting systems.

Then, a seven-step model for reporting based on a synthesis of the work done by Hambleton and Zenisky (2013) is introduced to present the report development process

in light of the literature in this area, and an updated checklist for designing reports is provided. Topics of interest in this section include some reflection on the interdisciplinary nature of reporting and what can be learned from other fields, the need for documentation such as user guides alongside reports, and the need for ongoing research on reporting. From there, the chapter concludes with attention given to challenges and opportunities in reporting, such as next-generation assessments, subscore reporting, technology-based reporting, advances in data visualization, and reporting in a formative context.

CONTEXT FOR REPORTING AND RESEARCH

> Over the years, I have been struck by the contradiction between the efforts and successes in producing sound technical assessments, drawing samples, administering the assessments, and analyzing the assessment data and the effort and success in disseminating the assessment results. (Hambleton, 2002, p. 193)[1]

The body of research on educational and psychological measurement largely came into being in the 20th century. During that time, most of the more well-known tests were, as an unofficial rule, both norm referenced and multiple choice, and the focus of the measurement field was largely on advancing the technical, behind-the-scenes statistics and procedures to develop tests and validate proposed uses (see Clauser, 2019, for a history of classical test theory; see also Clauser et al., this volume, for a history of educational measurement). Communication with users has not always been prioritized as a matter of theory *or* practice, nor was the compilation of evidence establishing the potential score interpretations as reliable and valid themselves, separate from scores.

In reflecting on the history of *Educational Measurement* as a volume chronicling key elements of this field, it is worth noting that results reporting only gained its own chapter in this edition. This is not to say that reporting has been deliberately or consciously neglected in the editions curated by Lindquist (1951), Thorndike (1971), Linn (1989), and Brennan (2006). It has been discussed briefly in prior editions: Indeed, for example, in the 1951 edition, Mosier's chapter on "Batteries and Profiles" (Mosier, 1951) provided some guidance on the elementary principles of profile construction as

> a graphic representation of a set of test scores for a single individual in which the tests are represented by ordinates spaced along the horizontal base line and the magnitude of each score is represented by plotting the point at the appropriate height on that ordinate. (p. 795)

In the 2006 edition, Cohen and Wollack noted that reporting is "clearly essential to test validity," but at the time it was "only just beginning to receive rigorous attention" (p. 382). Now, the time has come.

Reporting test results is a topic that cuts across psychometrics and many additional fields, including communications and marketing, graphic design, cognitive science, and information processing. In the same way that tests and testing practices have evolved, and stakeholder interest in and use of reports has changed along with how stakeholders

consume and use data more generally (Zenisky & Hambleton, 2012b), the way this profession has handled reporting has necessarily changed and is still evolving. No longer are reports simply data dumps with the goal to get as much information as possible onto a single page, and the literature on reporting is now much more substantial than even in the early 2000s.

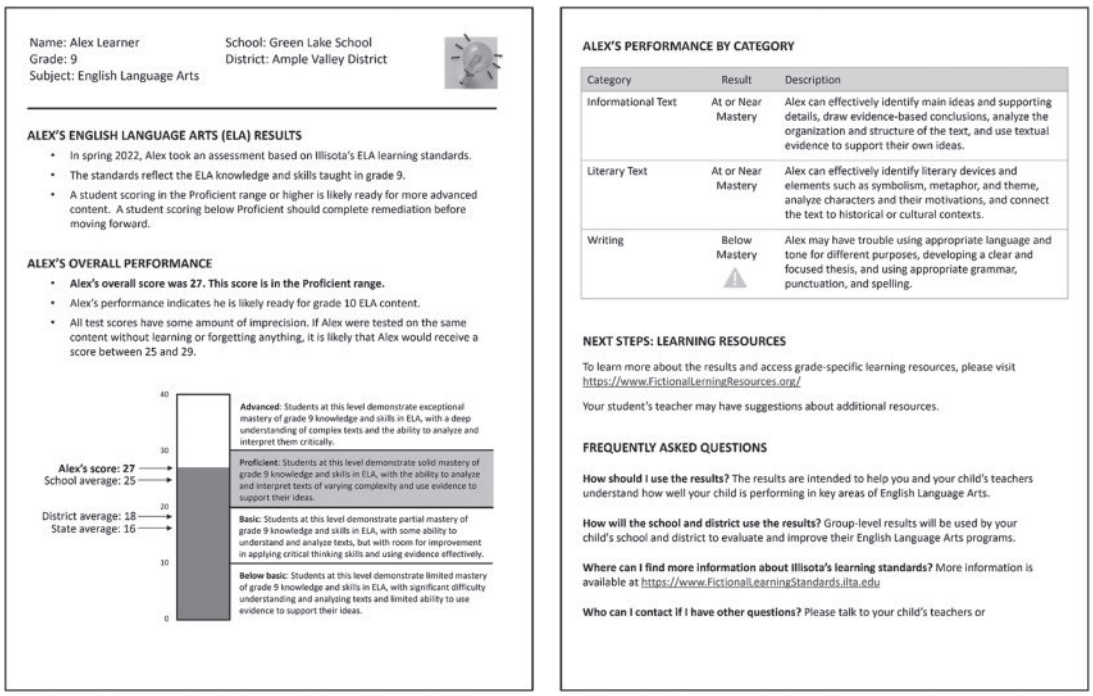
A number of historical events and movements have also contributed to greater awareness of the impact of results reports. Reflect for a moment on several of the tests, testing programs, and assessment-related initiatives that have appeared over time at the forefront of the public consciousness involving tests: the introduction of Army Alphas and Army Betas for military selection in the early 20th century (Waters, 1997); the rise of the SAT (formerly Scholastic Aptitude Test) with the objective to level the playing field for college admissions (and waves of pushback to the use of standardized tests for that purpose) (Lemann, 1999); the publication of *A Nation at Risk* during the Cold War (National Commission on Excellence in Education, 1983; the standards-based revolution and accompanying transition from norm- to criterion-referenced testing (Hamilton et al., 2009); the No Child Left Behind Act of 2002 in the United States and the accountability imperative of testing affecting most schools and its subsequent reauthorizations (Linn et al., 2002); the rising interest in national and international comparisons afforded by assessments such as the National Assessment of Educational Progress (NAEP), the Programme for International Student Assessment (PISA), the Progress in International Reading Literacy Study (PIRLS), and the Trends in International Mathematics and Science Study (TIMSS) (Fischman et al., 2019); and the evolving "opt-out" movement in many places (Kirylo, 2018). What is undeniable through reflection on these (selected) watershed moments in the history of assessment is the social impact of test results and how communication of results and the context for those results can both reflect and refract historical shifts in society. However, it is equally important to note that the quality of the reporting of results across these initiatives is variable: Typically, the data that are gathered lend themselves more easily to summative statements about proficiency on a relatively general or global level, and many reporting efforts falter when it comes to providing actionable guidance for intended users of such data, particularly when it comes to instructional uses.

Indeed, at present, stakeholders are often more skeptical consumers of data and seek additional information whenever possible, leading to something of a tension between what stakeholders want and what interpretations the psychometric models strictly support. In this way, testing agencies must thread a needle of reporting transparency, quality, and quantity that speaks to the very fundamentals of test development and validation where, it is our belief, results reporting should be among the very first considerations. Results reports are often the final and most public-facing aspects of any assessment system. But, rather than being left to something that is handled late in the development process, within the context of a principled approach to assessment design, reporting must be prioritized as part of assessment development and validity considerations to inform all the decisions that follow.

What Is a Results Report?

The topic of reporting begins easily enough with a paper document, often a single page, printed double-sided, that provides test results for one individual. Figure 13.1 provides an example based on a fictitious K–12 testing program. This kind of report is broadly representative of what an individual might receive after taking a college admissions test, a standardized educational assessment, a credentialing test, or a psychological test battery. The information communicated on the individual score report in each of those areas can be tailored to the recipient, but the commonality is an accounting of the performance of a single individual on a test taken. The actual results included can be represented in a wide range of ways, such as with numbers, text/labels, and graphics, and the primary aim of this document is to be informative at that individual level ("Here's how you performed"). In some testing contexts, paper reports have increasingly given way to electronic versions of the same information. These are sometimes accessed via email or, increasingly, a secure login to a reporting portal of some kind, but the reported information remains generally the same for individuals regardless of report delivery mode. Such reports can also be formatted for printing if that is an anticipated action users will take.

For some individual reports, it should be noted that the report not only serves the function of reporting status, but also offers prediction and/or guidance for next steps. This is perhaps most common in educational assessment, but is also found in reports



Name: Alex Learner
Grade: 9
Subject: English Language Arts

School: Green Lake School
District: Ample Valley District

ALEX'S ENGLISH LANGUAGE ARTS (ELA) RESULTS

- In spring 2022, Alex took an assessment based on Illisota's ELA learning standards.
- The standards reflect the ELA knowledge and skills taught in grade 9.
- A student scoring in the Proficient range or higher is likely ready for more advanced content. A student scoring below Proficient should complete remediation before moving forward.

ALEX'S OVERALL PERFORMANCE

- **Alex's overall score was 27. This score is in the Proficient range.**
- Alex's performance indicates he is likely ready for grade 10 ELA content.
- All test scores have some amount of imprecision. If Alex were tested on the same content without learning or forgetting anything, it is likely that Alex would receive a score between 25 and 29.

ALEX'S PERFORMANCE BY CATEGORY

Category	Result	Description
Informational Text	At or Near Mastery	Alex can effectively identify main ideas and supporting details, draw evidence-based conclusions, analyze the organization and structure of the text, and use textual evidence to support their own ideas.
Literary Text	At or Near Mastery	Alex can effectively identify literary devices and elements such as symbolism, metaphor, and theme, analyze characters and their motivations, and connect the text to historical or cultural contexts.
Writing	Below Mastery	Alex may have trouble using appropriate language and tone for different purposes, developing a clear and focused thesis, and using appropriate grammar, punctuation, and spelling.

NEXT STEPS: LEARNING RESOURCES

To learn more about the results and access grade-specific learning resources, please visit https://www.FictionalLerningResources.org/

Your student's teacher may have suggestions about additional resources.

FREQUENTLY ASKED QUESTIONS

How should I use the results? The results are intended to help you and your child's teachers understand how well your child is performing in key areas of English Language Arts.

How will the school and district use the results? Group-level results will be used by your child's school and district to evaluate and improve their English Language Arts programs.

Where can I find more information about Illisota's learning standards? More information is available at https://www.FictionalLearningStandards.ilta.edu

Who can I contact if I have other questions? Please talk to your child's teachers or

FIGURE 13.1
Simplified Sample of a K–12 Results Report

for other contexts where test users need help in identifying areas for improvement (for example, failing candidates in credentialing). This element of reporting can use subscore reporting techniques to identify areas of relative strength and weakness; in addition, rule-based/logic triggers (or, increasingly, artificial intelligence) can be implemented to evaluate response data and specify actionable next steps on a report. An example of this is a reading test that, as part of the reported results, provides a Lexile level to an individual and then suggests texts by name that are level appropriate (see Figure 13.2), or a mathematics assessment that reports relative weaknesses and structures reporting of such skills with links to specific online math lessons. Such connections, when included in reports, offer direct actions to the test taker and/or other users that take some of the guesswork out of what comes next, which is especially valuable when the report targets a test taker or families, who may well benefit from direct scaffolding in the area of "next steps."

Group reports are also an essential part of the reporting landscape, and the format and audiences for those are more varied than for individual reports. It is when results for groups of test takers start being aggregated that the extent of the *who* and *what* of group reporting comes into view, and the idea of what a "report" is starts to expand exponentially. Consider the *who* of reporting for a moment: In an educational/K–12 setting, a group could, for example, consist of:

- a subset of students in a classroom, selected purposefully by a teacher or instructional administrator;
- all students in a classroom;
- selected students within a grade level grouped by some socioeconomic or demographic quality;
- all students in a grade level within one school in one district;
- all students in a grade level in multiple schools within a district;
- all students in a grade level within a state or territory; and/or
- all students in a grade level within a country.

These various groupings are not of equal interest to all users of assessment results. What a classroom teacher will focus on in a group report is necessarily quite different from what a school or district administrator will look at, which is different again from what

Your Reading Level: 1340L **Current Reading Range: 1240L – 1390L**	The best way to improve your reading level is to read books in your range. Consider the suggestions below.	
Book	**Author**	**Level**
Hope For Animals and Their World	Jane Goodall	1240L
Stories Well Told: Science Fiction	Valerie Bodden	1340L
Core Four	Phil Pepe	1390L

FIGURE 13.2
Sample Report Section With Lexile-Based Book Suggestions

administrators in a state or provincial office of education consider most important. Although the idea of group reports is simple in that a group report is most basically some kind of aggregation of individual performance results, it broadly encompasses many different groups of individuals, data about their performance, and levels of communicating results.

Similar kinds of conceptualizations of what is a "group" for reporting can be applied to other testing settings. For example, in credentialing, a group could be all test takers testing at a certain test center within a defined testing window, all test takers from a certain institution or training school within a year, or all test takers grouped by a socioeconomic or demographic quality within a year—these are dependent, again, on the audience and the planned use of the data.

Because of the many possible group selections for aggregate reports, group reporting is increasingly occurring (or at least originating) as a function of database queries rather than static report documents. To a point, dependent on the user group and the intended use(s) of group-level data, the process of accessing aggregate test results involves using interactive, web-based tools that tend to reflect a continuum of analysis (Zenisky & Hambleton, 2012b). This continuum spans from tools that provide results that are purely descriptive in nature to those that function more akin to a statistical analysis package to carry out original analyses, including significance testing. Examples of the former include the data tools for many U.S. states, such as Massachusetts (http://profiles.doe.mass.edu/statereport/mcas.aspx), Florida (https://edstats.fldoe.org/SASPortal/main.do), and Texas (http://texasassessment.com/administrators/). The International Data Explorer tool is the premiere example of the latter class of online assessment results tools because it is a "one-stop shop" for access to results from PISA, PIRLS, TIMSS, and the Programme for the International Assessment of Adult Competencies (https://nces.ed.gov/surveys/international/ide/). This portal, which mirrors that of the NAEP Data Explorer (https://nces.ed.gov/nationsreportcard/data/), permits users to run simple or complex descriptive results and carry out analyses of performance data with respect to innumerable additional variables gathered through participant and educator surveys. Results can be run to produce tables as well as highly customizable graphics.

Conceptualizing Reports and Reporting

> Know the communicative purpose of the display and do not try to do too much. (Wainer et al., 1999, p. 304)

A critical aspect of reporting that must be raised at the outset of this discussion concerns the idea of *effectiveness*. Much of the research and discussion that focuses on results reporting aims to support the development of "good" reports, but the quality of a report is a concept that deserves some reflection. What *is* an effective report? Ryan (2006) established a report itself as a form of communication, and a starting place to answer this question can be traced back to several key advances in the results reporting literature. First, the work of Wainer in focusing on visual displays of quantitative data (e.g., Wainer, 1992, 1997a, 2005, 2009; Wainer et al., 1999) is especially

instructive in that it draws across disciplines to formulate broad principles that apply to individual and group reports alike, such as creating visuals that are high in information and low in adornment, labeling clearly and fully, and using spacing to aid perception.

This line of research advanced by Wainer and others calls on those responsible for developing reports to think critically about the intended story of each data display as well as the nature of the data being presented. There are numerous examples in Wainer's writings of cases where graphics (assessment related and otherwise) were revised to show a very different story, one that was obscured by the original data presentation. Building on Wainer's work on graphical presentation and its considerable implications for results reporting of assessment data, then, effective—in part—draws on adhering to basic principles of communication and cognitive processing to create reports that are purposefully crafted to communicate specific information.

The foundational work on quality reporting done by Jaeger (Jaeger, 2003; Jaeger et al., 1993) likewise laid the groundwork for differentiation of reporting by audiences. By referring to "NAEP's audiences" (in the context of the NAEP) and reframing school report cards with the intended recipient in mind, reporting has effectively had to shift from one size fits all to a model that acknowledges and responds to the informational needs and interests of users of test data. This is a perspective that has rightly permeated current thinking about reporting and provides a conceptual basis for reporting efforts that are informed by research with various articulated user groups.

Figure 13.3 provides an updated perspective on Jaeger's (2003) nine tables offering a road map for audience-specific research activities for report development. The premise of Jaeger's (2003) work was that there are some key questions that should be asked of each stakeholder group, to learn not only what is of interest to them, but also how such information could and should be provided to them. And, with each stakeholder group, there are different methodologies available to report developers for the purpose of investigating those key questions, and those methods, naturally, provide quantitatively and qualitatively different data. Figure 13.3 links the main considerations in the design of research on reporting: For agencies that engage in these questions, it may be helpful to frame the work to be done relative to these considerations and build out the work relative to the priorities and available resources of a testing program.

Indeed, Jaeger's (2003) contributions, including the differentiation of users (and the accompanying need to probe and understand reporting interests of different groups), provide a direct line to another key advance in conceptualizing quality reports, which is the presentation of general models for report development and evaluation (Hambleton & Zenisky, 2013; Zapata-Rivera, 2011; Zenisky & Hambleton, 2012a, 2015). With regard to model-based approaches to report development, the focus on audiences and users as something to be articulated and respected propagated a shift in the process of report development, where Hambleton and others suggested that reports should evolve through audience-specific research and development, and reporting aims are discussed and user input is solicited at multiple points, to inform report development from conceptualization to implementation and maintenance. This advance helps to

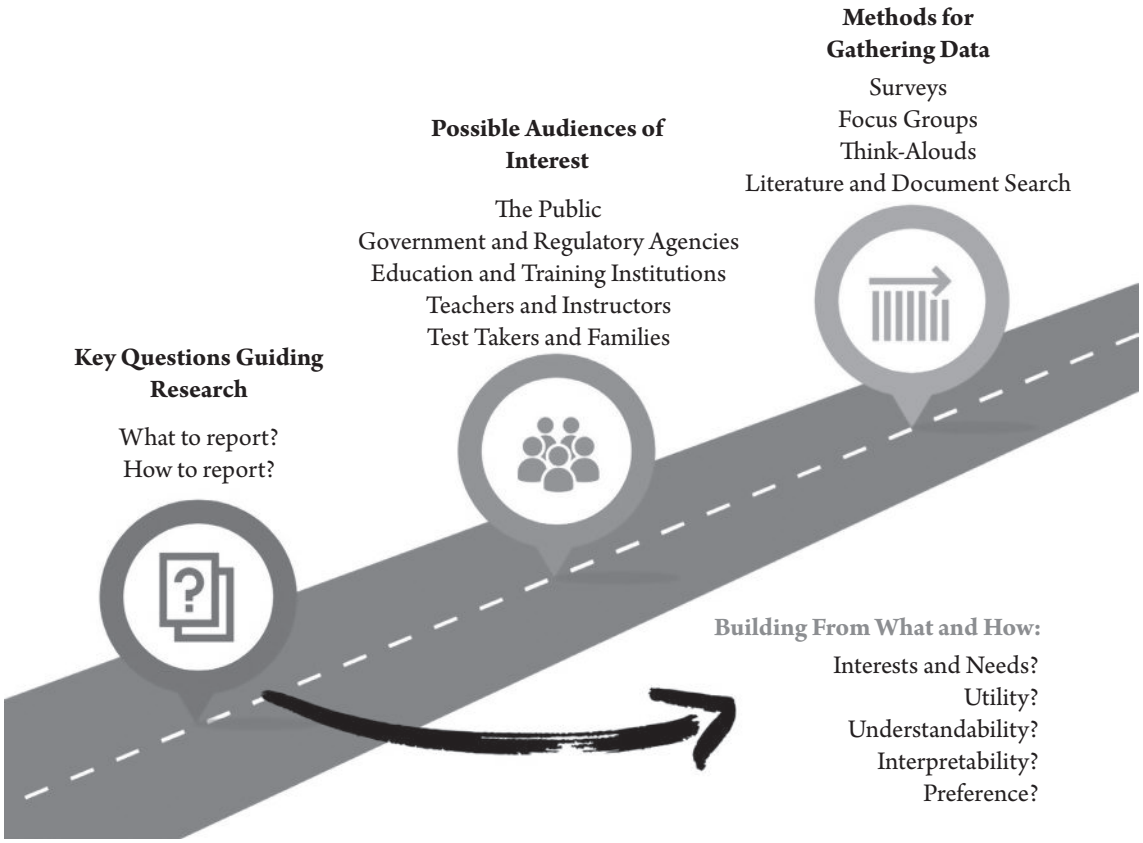


FIGURE 13.3
A Road Map for Audience-Based Research

Note. Adapted from *NAEP Validity Studies: Reporting the Results of the National Assessment of Educational Progress* (Working Paper 2003-11), by R. M. Jaeger, 2003, U.S. Department of Education, Institute of Education Sciences.

ensure that a key dimension of how report quality is defined is informed by direct communication with and solicitation of input from the intended users.

Another fundamental source for informing a working definition of an *effective* report is drawn from Hattie's approach working with educators, under the umbrella of his "Visible Learning" initiative (Hattie, 2009, 2010). In the realm of results reporting, much of Hattie's work has sought to elevate the importance of the audience and supporting stakeholders in using assessment data. The lessons of his research on learning, achievement, and results reporting transcend their original context to apply to all of the reporting contexts relevant here. In Hattie (2010), 15 principles for establishing the validity of results reports are enumerated, grouped in topics including validity of reports, sources of validity evidence for reports, and design principles for reports. Among these principles are the ideas that reports should have a specific theme and be designed to address specific questions posed by the stakeholder and that, from a conceptual point of view, reports should be conceived of as actions, not screens to print. These are critical principles that, similar to the ideas in Wainer's work, speak to the notion of what effective is that will guide this chapter.

Reporting and Professional Standards

The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 2014), offers an additional lens for evaluating report effectiveness for the contexts described in Figure 13.4: educational assessment, workplace testing and credentialing, and psychological assessment. The *Standards* are presented in clusters, and two clusters specifically address results reporting: one focusing on reporting and interpretation and another on test takers' rights to fair and accurate reports. The first standard from the first cluster is the most emblematic of the kind of recommendations provided:

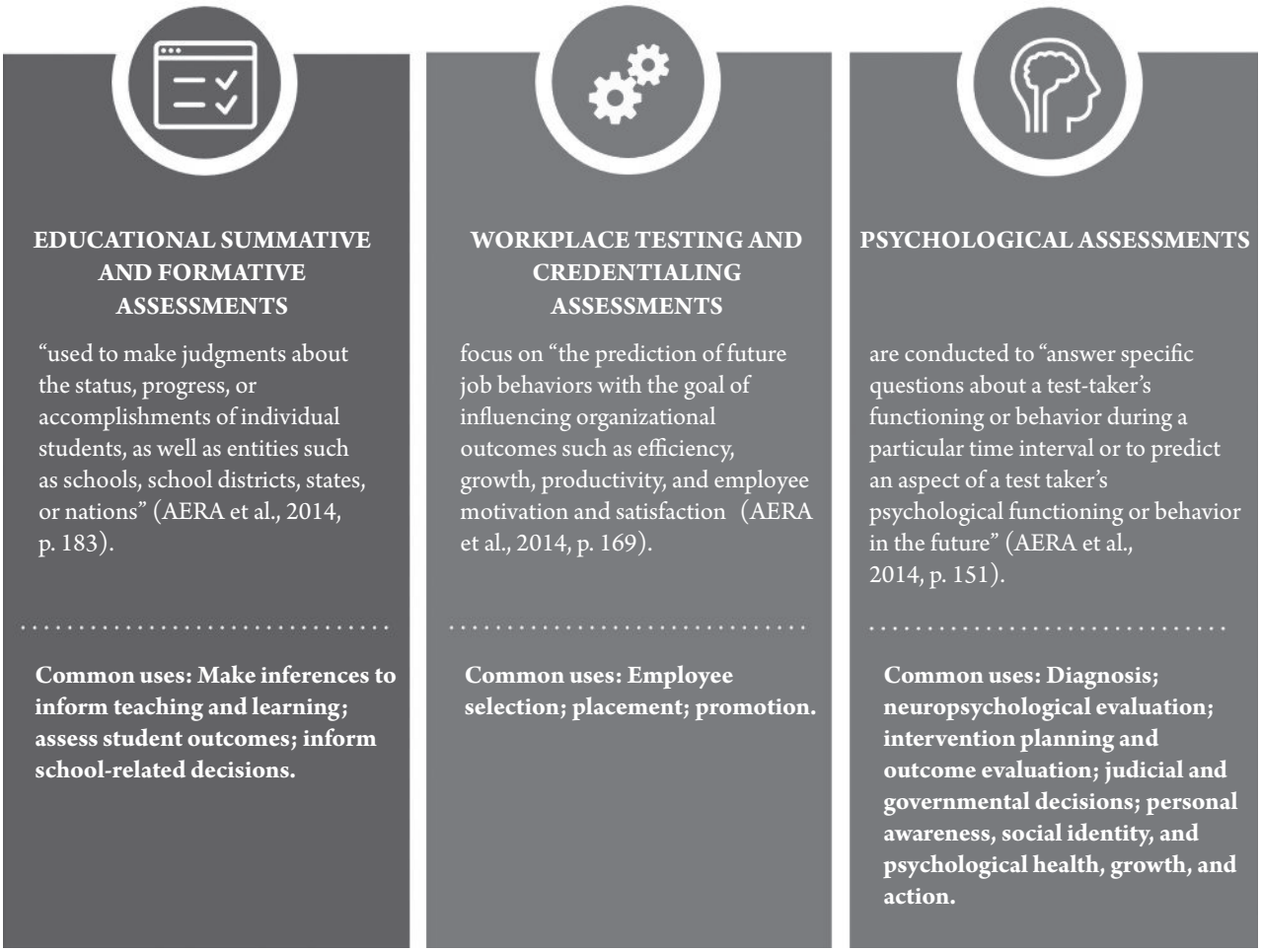


FIGURE 13.4

Overview of Assessment Contexts for Reporting, per the *Standards for Educational and Psychological Testing*

Note. Information is drawn from *Standards for Educational and Psychological Testing*, American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014.

> When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used. (AERA et al., 2014, p. 119)

With sentences such as "provide interpretations appropriate to the audience" in the text of the standard and "reports and feedback should be designed to . . . minimize potential negative consequences" in the accompanying commentary, the *Standards* offers reporting ideals in ways that are necessarily broad to account for guidance that spans highly varied testing purposes. Examples of how those ideals might be achieved are discussed, but the specific approaches in the examples are options rather than requirements, in alignment with the thinking that there is no one-size-fits-all report design that will succeed in every testing context.

Within the two clusters, there are standards addressing the need to ensure the validity of automatically generated interpretive text (Standard 6.11) and select performance labels that support intended inferences without being stigmatizing (Standard 8.7; e.g., O'Donnell & Sireci, 2021). There are numerous performance-level labels in use, as illustrated by the word cloud in Figure 13.5 based on O'Donnell's (2020) review of labeling practices for statewide assessments in the United States (larger text represents more frequent use of a word). There are also standards for more procedural aspects of reporting, including report delivery and handling material errors. Beyond the standards in those clusters, O'Donnell and Zenisky (2020) identified 36 standards (AERA et al., 2014) that apply to results reporting, many of which are mentioned later in this chapter along with standards from the National Commission for Certifying Agencies (2014), which are similar but geared toward credentialing.

FIGURE 13.5
Performance-Label Word Cloud

The International Test Commission's Guidelines on Test Use (International Test Commission, 2013) are also instructive here: While some of those guidelines offer assistance in the selection of tests and test administration, Section 2.7 focuses on interpretation with 12 substandards pertaining to generalization of scores, reliability and validity of results and implications for meaning, and minimization of bias and social stereotyping that may negatively impact individuals and groups. Section 2.8 of these ITC guidelines further relate to communication of results, addressing privacy concerns and providing guidance about communication strategies that take audience interests into account. Overall, the guidelines from AERA et al. (2014) and other groups place validity and understanding one's audience front and center, promoting the idea that an effective report is one that guides users toward valid interpretation and uses, leaving it up to developers to find the exact content and design features that will be most successful in that task.

From Score Reporting to Results Reporting

Historically, the topic of communicating test results has been referred to as score reporting and, indeed, scores (primarily overall scale scores) were the main or only data contained on reports for many years. Over time, however, there has been a clear shift in most large-scale testing programs where what is being communicated to audiences is substantively different. Much of what is found in present-day reports extends well beyond simple scores to include achievement levels (e.g., *Pass, Not Yet Meeting Expectations, Proficient*), information to support comparisons to relevant groups (e.g., percentile ranks, reference group descriptive statistics), and numerical or categorical subdomain feedback. Other results sometimes included are item-level performance data, where item-level results can be provided for individual test takers or groups of test takers, and growth results (Zenisky et al., 2019), which use test scores from previous administrations to project test performance in the future.

Although it is certainly true that most of the data presented on current reports are scores or derivatives of scores, the shift in language being advanced here—from "score report" to "results report"—supports a more inclusive perspective. This change in emphasis is intentional in that it reinforces a more expansive approach to representing test performance. The task explicit in this shift is to consider and explicitly prioritize ways to communicate test information that are richer and more engaging. For example, there is growing research on interactive reporting tools such as dashboards, discussed later in this chapter.

Our recommendation, reflected in the terminology shift, is to approach the task of reporting assessment results as sharing multiple pieces of information with unique value rather than treating data other than the overall score as secondary in importance. This line of thinking may help develop reports that not only fulfill informational purposes and contribute more meaningfully to reporting systems, but also guide users to what comes next on the basis of the results, aligning with the third of Hattie and Timperley's (2007) reporting questions: Where to next?

PERSPECTIVES ON REPORTING FRAMES AND VALIDITY

> Clear and useful score reports support users in making appropriate score inferences and have an important role to play as part of efforts to explain the validity argument for a test to key stakeholders. (Hambleton & Zenisky, 2013, p. 479)

Reporting as Validity in Action

Validity is the most essential consideration in the process of developing and evaluating tests. In turn, how test results are interpreted and used is central to how validity is defined in the AERA et al. (2014) *Standards* because validity there is understood as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). Interpretations and uses are also key to the argument-based approach to validation (Kane, 2006, 2015), which provides a practical framework aligned with the *Standards*. Following this approach, at the outset of test development, test developers must first create an interpretation and use argument (IUA), listing all proposed interpretations and uses along with the reasoning for each, and then proceed to a validity argument, which involves a strategic evaluation of the claims and assumptions in the IUA. The development of an IUA, then, can provide a clear evidentiary basis for reporting (Ferrara & Lai, 2016).

Similarly, based on the *Standards* (AERA et al., 2014), test developers are asked to articulate all intended interpretations of results for specified uses (Standards 1.0, 1.1, 11.1) and later gather and evaluate validity evidence for each of those interpretations (1.0, 1.2, 1.11–1.25, 11.1). Intended interpretations and uses are mentioned in standards regarding scores (5.1, 5.4), precision (2.0), norms (12.5), and various other topics. A smaller number of standards focus on ensuring that those intended interpretations come to fruition when reports reach their audience:

> Standard 3.1: Those responsible for test development, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant sub-groups in the intended population. (p. 63)

> Standard 13.5: Those responsible for the development and use of tests for evaluation or accountability purposes should take steps to promote accurate interpretations and appropriate uses for all groups for which results will be applied. (p. 211)

This notion of effective reporting is in line with a view of validity heavily informed by the work of Messick (1989), with an emphasis on consequences and use. Reporting is essential to validity because it is the means through which intended interpretations are realized, and this is critical to interpretation, particularly in terms of the intended (and unintended) consequences of testing (per Messick, 1989). The greater focus on *intended* rather than *actual* interpretations in the AERA et al. (2014) *Standards* is likely because the former are needed to start building an IUA and are available earlier in the test development process. However, the interpretations made by users on the basis

of reports should also be a major component of validity evidence. If a report fails to guide users toward the validated interpretations for an assessment, the chances that the assessment program will meet its goals are greatly diminished, regardless of how much evidence exists in support of ideal, but unrealized, interpretations. Ryan (2006) also espoused this connection between validity and interpretation, suggesting that elements or features of reports that foster or lead to unsupported interpretations have a corrupting effect on validity. (See also Lane & Marion, this volume.)

Tannenbaum (2019) described this issue in terms of alignment: "A score report that is not well aligned with the test is of little value; similarly, a score report that is well-aligned, but not communicated to users in a way understandable to them is of little value" (p. 9). This is a critical point that speaks to an important dimension of effective, in that a report without alignment or meaning is essentially devoid of consequence and occupies space without adding value. To ensure alignment, Hambleton and Zenisky (2013) proposed explicitly linking validity evidence and the purpose of a test to the contents of results reports and subsequently investigating the characteristics and information needs of the intended audience to determine how to best present that content (all steps of the Hambleton and Zenisky model are described in a later section). A common thread among Tannenbaum (2019) and Hambleton and Zenisky (2013) is the view that effective reporting is critical to validity, and ensuring that reports can be interpreted as intended to fulfill specific purposes is as important as having other sources of validity evidence.

Several authors have proposed alternate notions to better situate reporting as part of the validation process, recognizing its critical role. These include the following:

- The validity of reports: the notion that reports deserve their own validity arguments (Hattie, 2010);
- User validity: the perspective that the conceptualization of validity from the AERA et al. (2014) *Standards* should be expanded to include a user-centered source of evidence based on the accuracy and effectiveness of interpretations based on test output (MacIver et al., 2014);
- Report interpretability: the view that the interpretability of reports should be considered an aspect of validity (Van der Kleij et al., 2014); and
- IUA source materials: the view that reports and other technical documentation of tests are source material for the IUA for an assessment (Ferrara & Lai, 2016).

O'Leary et al. (2017) expanded on those ideas, proposing an approach that aligns with the unified conceptualization of validity (Messick, 1989). Instead of discussing new types of validity, O'Leary et al. (2017) recommended expanding the five sources of validity evidence—content, response processes, internal structure, relations to other variables, and consequences—to include interpretability: evidence focusing on the "adequacy, accuracy, and effectiveness of user understanding of scores and the consequences of testing" (p. 20). Further, they proposed that evidence based on interpretability and evidence based on consequences be treated as having equal importance as

the four other sources, which traditionally have only yielded evidence in support of possible interpretations.

Alternatively, the five current sources of validity evidence could be adapted to encompass results reports more explicitly. Tannenbaum (2019) provided several ideas of how this might happen. Evidence based on test content, which is usually about the relationship between the content of the test and its target construct, could also include evidence that the content of the report is well aligned with the content of the test. Evidence based on response processes, which focuses on how well the test elicits the strategies or cognitive processes that are key to the construct, could incorporate evidence that users visually navigate the report as intended, spending more time on the areas the design of the report was meant to highlight.

Evidence based on internal structure is typically about how well the relationship among items or portions of a test reflect its target construct and the extent to which those relationships stay consistent across subgroups of test takers. When applied to reporting, this source could involve gathering evidence of how well report users understand the ways in which different parts of the report relate to one another (e.g., how a sentence about precision should inform conclusions based on a graph with content area results). It would also be relevant to evaluate how different groups within the target audience interpret and use the information reported (e.g., groups from different socioeconomic backgrounds, test takers scoring below and above a passing score).

Gathering evidence based on relations to other variables in the context of evaluating results reports is more complex. Traditionally, this type of evidence comes from analyzing the relation between test scores and other measures. To evaluate reports, Tannenbaum (2019) suggested investigating the match between students' competency as communicated on their score report and as evaluated by teachers, noting that convergence would be confirmatory evidence and discrepancy would be more challenging to interpret (it could reflect differences in what was tested compared to what teachers considered in their evaluation).

Finally, validity evidence based on consequences of testing has a strong connection to reporting. The interpretations individuals make and any actions they take in response to results reports are a direct consequence of testing. So are potential misinterpretations, inaction, and undesired actions. This type of evidence, which could employ multiple methods, should shed light on the extent to which reports are interpreted and used as intended based on reasonable expectations or the IUA for a test.

Whether testing organizations plan test development and validity research in alignment with the argument-based approach to validity or a framework such as principled assessment design, described next, effective reporting should be treated as essential to validity and given due consideration.

Reporting and Principled Assessment Design

Principled assessment design (PAD) refers to several approaches, including evidence-centered design (Mislevy & Haertel, 2006) and assessment engineering

(Luecht, 2013), intended to offer a more integrated pathway for designing or improving assessments and building validity arguments. Huff et al. (this volume) offer a comprehensive overview of the framework, detailing how—in contrast to more conventional methods that may lead to siloed activities—it employs strategies to ensure coherence "from construct definition through task development and score inference" (p. 447).

Both PAD and the argument-based approach to validity reflected in the AERA et al. (2014) *Standards* rely on the explicit articulation of ideas to guide the design and validation process. However, while the argument-based approach calls for the articulation of proposed interpretations and uses of assessment results (along with their rationales), PAD requires integrated documentation and articulation of the assumptions and decisions related to *all* aspects of test design and validity research. This focus on approaching the assessment cycle in a more unified way is ideal for reporting work, which in the past seemed to be treated as an afterthought in the test design process (Katz, 2019).

PAD is a promising framework to help those involved in test development meet present-day challenges, including greater distrust and scrutiny of assessments and the increasingly common expectation that one assessment may serve multiple purposes (Huff et al., this volume). In practice, approaching results reporting in the context of PAD may take many forms. It may involve creating a prospective score report early on to promote discussions about what information the assessment should provide (Slater et al., 2019; Zapata-Rivera et al., 2012), using the prospective score report to clarify the claims to be made about test takers (Zieky, 2014), and changing test design specifications to align with the desire to report specific information, such as subscores (Sinharay et al., 2019).

What is common across these examples, and central to PAD, is the idea that reporting considerations are a key component of assessment design—not a step to be taken when other important test design decisions are already finalized. In such an unfortunate scenario, the report must conform to those established test design decisions, which may not be aligned with the information needs of report users. PAD offers an approach for considering test content, format, and reporting in tandem, making it more likely that results reports will be effective in promoting validated inferences. This idea of articulating reports and their purposes in advance might be viewed as backward design, in the sense of starting with the product and working through the process early on to get to that place in the end, but such an approach is quite forward thinking and in effect prioritizes understanding and actions associated with reports before it is too late to change course.

Reports as Data Stories

One of the ways of making statistical results more meaningful to intended audiences is to report the results by connecting them to numbers that may be better understood than test scores and test score scales. For example, to relieve the concern many persons had about flying after the TWA crash a few years ago,

> the airlines reported that there is a single plane crash every 2,000,000 flights. In case the safety of air travel was still not clear, the airlines reported that a person could expect to fly every day for the next 700 years without an accident. Probably some people felt more confident after hearing these statistics reported in this way. Knowing that the probability of being in a plane crash is less than .0000005 may not be so meaningful. (Hambleton, 2002, p. 194)

Building on the approaches to validity and PAD referenced here, a key idea to raise at this point is that of the *data story*, where results reporting can be conceptualized as a coherent and planned approach to communication about assessment data, rooted not only in the validity of the inferences being made but also in the validity of the communication about the data. Data story is a term that has emerged from business and marketing settings in recent years to describe a highly coordinated strategy for talking about data in various nonassessment contexts, but, given the nature of data in reporting, there is a natural relevance to this topic in educational and psychological testing. The basic concept of a data story is that it is a thoughtful and intentional approach to sharing data that brings together data visualizations and compelling narratives, targeted to specific audiences, that aims to help intended users of the data understand the data and take action where appropriate (Hooper, 2021). It is not simply a matter of better (or different) graphics, but rather an orientation on the part of report developers to identify a story to be told with data and then use tools such as visualizations and one or more narrative patterns to explain something about the data (and why the data matters), in the context of a specific vehicle for communication (e.g., a results report, a presentation, or an online reporting tool).

Bach et al. (2018) defined a narrative—in this data story context—as giving shape to the *events* in a data story, following a specific narrative pattern. The pattern(s) that underlie any data story can vary from context to context, and the choices made are predicated on a narrative's intent. Narrative intent is an important idea here, in that it reinforces the purpose of the communication effort and can be linked to the context of assessment through articulation of both (a) test purpose and (b) report purpose. Examples of intents that might underlie data stories from the data story literature could include enlightening audiences, evoking emotional responses, spurring action, and questioning beliefs and behaviors. It is not hard to move from these general intentions to the aims of results reporting with respect to these intents.

To select and use any narrative pattern, the author must have a story idea, an idea of the intended audience, and an intended effect such as sympathy, action, information, or explanation (Bach et al., 2018). Conceptually, this links quite closely to the layers of reporting put forth by Behrens et al. (2013), in which three layers of reporting were articulated. Layer 1 corresponds to information communication (in effect, the "what" of reporting, asking, What is the story to be told?), while Layer 2 is couched in social activity (How are people approaching the story and who is the audience for the story?). The third layer of Behrens et al.'s (2013) approach aligns to societal transformation

(asking, What is the moral of the story?)—in effect, the intended outcome, or the *why* of the act of data communication.

Drawing on visualization and data story work, as well as arts and literature, there are several ways to think about narrative patterns. One overarching narrative strategy that has been identified in this area is explanatory versus exploratory orientations (Thudt et al., 2018). This has been discussed in the context of results reporting by Zenisky and Hambleton (2012b) relative to digital reporting efforts. Explanatory reporting efforts are predicated on the report developer's conscious and explicit choices about content, appearance, and interpretation support, while exploratory efforts typically align to interactive tools that offer users flexibility and personalization of the experience of the data story. Most traditional results reports in use for educational assessments in the early 21st century follow the explanatory narrative strategy, while publicly available anonymized databases on state websites draw on the exploratory model, with drop-downs and selection boxes that allow for customization.

Another narrative pattern to consider in the context of reporting involves prediction and how users might be presented with formative results to guide future action. Prediction, as a reporting aim, draws on the past to identify next steps. Thus, reports that address this aim may use data that might be presented in a static way using an explanatory strategy to set the stage for the results and include an exploratory component that offers the user the opportunity to navigate through to identify areas of strength and weakness and connect the results to specific actions.

At a similarly straightforward level, another approach to narrative patterns is to think of some narratives as linear (following a highly temporal or sequential route, such as the Harry Potter books tracing Harry's path from preadolescence to adulthood in the wizarding world) while others are nonlinear (such as the television show *Lost*, with the story told using flashforwards, flashbacks, and even sideways/parallel paths; or *The Godfather Part II* film, which juxtaposed events in the lives of the Corleone patriarch and his son to unfold both stories semisimultaneously). In the context of results reports, the analogue to a linear narrative pattern is the body of work that is produced to communicate the long-term trend results for NAEP (National Center for Educational Statistics, 2013). The overarching data story in long-term trend results is based on data that are linear in nature, with many visualizations used that aim to illustrate the performance of test takers over time.

Exploratory or explanatory and linear or nonlinear are not, however, the only narrative patterns that can be used for data stories. Bach et al. (2018) identified five different groupings of narrative patterns based on broad intent. These groupings are provided in Figure 13.6.

In reflecting on these groupings and the specific patterns within each one, it is important to note that any data story may well use more than one of these techniques to accomplish different intentions within the same document. For example, on a typical individual student results report for K–12 assessment in the United States, narrative patterns that can be spotted relatively quickly might include familiarization

Patterns for argumentation
These serve the intent of persuading and convincing the audience. Examples of these patterns are compare, concretize, and repetition.

Patterns for flow
These patterns serve the intent of structuring a sequence of a message or argument, setting the order, rhythm, and pace of the story. Examples here are reveal, slowing down, and speeding up.

Patterns for framing the narrative
Patterns here serve the intent of controlling how facts and data are perceived and understood. Examples of framing narrative patterns are familiar setting, make-a-guess, defamiliarization to challenge expectations/convention breaking, silent data, and physical metaphors.

Patterns for empathy and emotion
These serve to engage the audience with the content of the data story. There are overlapping narrative patterns here (reveal, slowing down, speedup, and concretize). There are also some novel patterns as well, such as breaking the fourth wall, humans behind the dots, and familiarize.

Patterns for engagement
This category serves the intent of connecting the audience with a story, to make them feel part of it and perhaps even offer them some control. Specific tracks that can be taken here include the use of rhetorical questions, call-to-action, make-a-guess, and exploration.

FIGURE 13.6
Narrative Patterns for Reporting Data Stories

Note. Adapted from "Narrative Design Patterns for Data-Driven Storytelling," by B. Bach, D. Stefaner, J. Boy, S. Drucker, L. Bartram, J. Wood, P. Ciuccarelli, Y. Engelhardt, U. Köppen, and B. Tversky, 2018, in N. Riche, C. Hurter, N. Diakopoulos, and S. Carpendale (Eds.), *Data-Driven Storytelling* (pp. 107–133). CRC Press. https://doi.org/10.1201/9781315281575-5

(using the student's name in personalizing the report), compare (to compare the student's observed score to a class, school, district, and state), and physical metaphors (with stoplight displays or pie charts to illustrate mastery of subdomain skills). This is quite common: Bach et al. (2018) commented on a number of use cases where multiple narrative patterns are used in data stories, pulling examples from different media sources.

The literature on results reporting itself offers a number of additional narrative patterns. One such approach is *question and answer*, where the flow of the report document is sectioned by a series of questions formulated by the report authors to reflect common questions from intended users, with answers provided. Another approach is reflected in the traditional individual results report that presents a high-level overall result, followed by a progressive "drill-down" to more granular series of data points,. In this case, the pattern of the data story moves in a linear fashion on the basis of the granularity of the data (from least to most), typically drawing in some comparisons and making some inferences highly concrete. The report might then conclude with a call to action (in a section entitled "Next Steps"). Figure 13.7 highlights the narrative patterns used in the sample report presented at the beginning of the chapter.

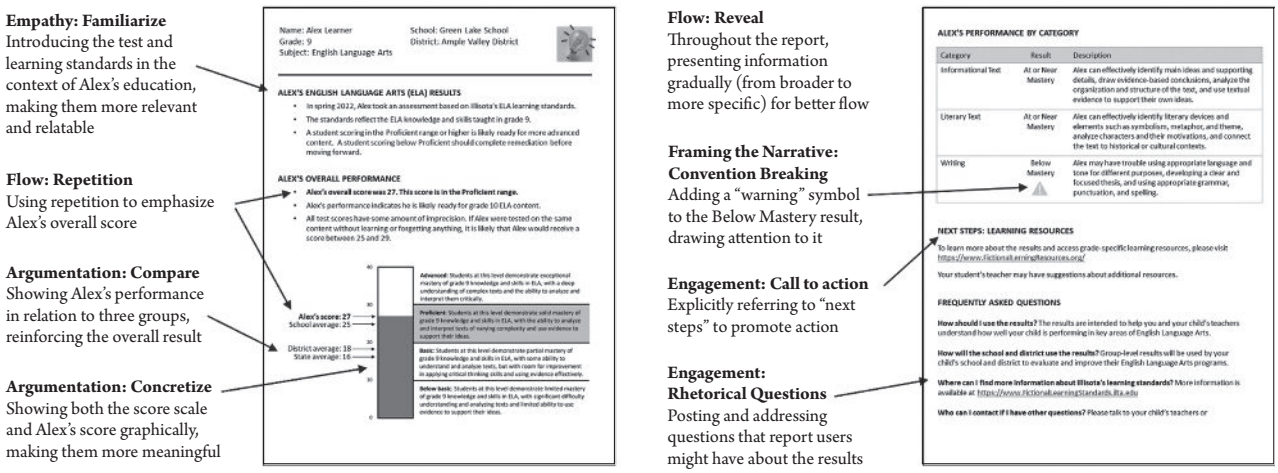


FIGURE 13.7
Simplified Sample Report With Annotations Highlighting Narrative Patterns

Hattie's (2010) work on the validity of reports likewise offered some guidance that aligns well with the idea of results reports as data stories. Principle 3 ("Readers of reports need a guarantee of safe passage") and Principle 4 ("Readers of reports need a guarantee of destination recovery") align to the underlying concepts of Bach et al.'s (2018) patterns for flow. Principle 5 ("Maximize interpretations and minimize the use of numbers") is a way into patterns for argumentation, and critically, Principle 8 ("Each report needs to have a theme") brings us full circle on the very idea of data stories and intent: defining the aim of the report and articulating the structural and communication techniques that can be employed to achieve the intent of a specific data story, for specific user groups.

The value of the data story and narrative pattern framework lies in the premise that reporting in educational and psychological testing is an intentional act of communication about data, and the approaches to conceptualizing and organizing the stories discussed here, drawn from outside measurement, have significant interdisciplinary value. Across assessment contexts and reporting settings, there are many data stories to be told (at both individual and group reporting levels) and many ways to accomplish the communication of assessment data. Under this paradigm, report developers must look at the data story of the report as a coherent whole, as well as how distinct report elements contribute to the story in a purposeful and organized way.

ESTABLISHING REPORTS, REPORTING SYSTEMS, AND REPORTING CONTEXTS

There is no one-size-fits-all design for what constitutes a "good" score report. Tests and test purposes are different, and users' data needs vary. No one magical visual display will make test results understandable to all users, and—sadly—there

> is no one perfect line of text that illuminates what standard errors are and why they matter. (Zenisky & Hambleton, 2015, p. 586)

In considering the breadth and depth of what reporting is more fully, it may be helpful to extend the metaphor of the data story to assist in defining critical reporting dimensions and aspects. Specifically, the questions of *who*, *what*, *where*, *when*, *why*, and *how*, long associated with journalism, can serve as something of an organizing framework for conceptualizing and disseminating results reports broadly in the context of assessment, in line with the data story metaphor.

Who?

The first of these organizing questions is *who*. In results reporting, there are in effect two very significant ways to address this fundamental point in report development. One of these involves the subject(s) of the report, in terms of whose results are being reported. Quite simply, a report can be crafted to describe the performance of one (a single individual) or many (more than one). From the perspective of developing a report, this first aspect of *who* concerns the unit of analysis for the report—individuals or groups (and how group membership is operationalized for the purpose of the report).

From that natural division arises the second way to address the *who* of reporting, which is the intended user. Reports are developed to describe the performance of an individual or a group on a specific assessment, and building on that, individual and group reports are typically developed for different intended audiences (to fulfill different reporting aims—the *why* of reporting).

The most detailed reports for individual test takers are typically disseminated to the test takers themselves and, depending on the testing context, others close to them in the social space, such as family members and educators. Indeed, reports for individuals are typically more relevant for test takers and their own families and for the test takers' instructors. In the context of credentialing, individual reports are provided to the test takers themselves, and the considerations of reporting are differently nuanced between the personal and the professional uses (O'Donnell & Sireci, 2019).

As the social distance from the test taker grows, the nature of the substantial detail around the performance of the individual can change. In the example of certification and licensure, individual results at the global level of pass or fail may be transmitted to a local or national governing body for record keeping and maintenance of professional standards, but the focus in this group-level report to this user group is not instructional or formative, and as such, subdomain results or other score breakdowns may not be of primary interest for these particular users. (Detailed results exploring comparisons between test-taker subgroups will be of interest, but such data are typically summarized and anonymized for governing body uses [record keeping, monitoring, and public communication].)

To that end, results for groups of test takers can be constructed to reflect groups of hyperlocal interest (such as a class, grade level, or school) all the way to aggregations

for national and/or international comparisons. School and district personnel may use group-level reports to identify academic strengths and weaknesses within classes, schools, and districts, and at the larger scale (such as states) statistical power can be used to help identify broader trends in student performance over time (Zenisky & Hambleton, 2015). Reporting units of interest can be based on geography (such as class, school, district, state) or other relevant demographic variables (e.g., race/ethnicity, language status, individualized educational plan status). Thus, results reports for groups can be reported on the basis of groupings constructed to reflect geography, demographics, and/or other dimensions of interest. The users of various group reports can be close (as in educators and coaches) or more removed (such as policy makers, the general public, and the media). The more removed such users are, the larger the aggregation (typically), and the purpose becomes more policy oriented in nature.

Though sometimes it is the individual score report that garners much of the interest in terms of reporting research, a great deal of foundational work on reports and reporting quality occurred within the context of evaluating NAEP group-level results (Zenisky et al., 2016), where research focused on interpretation of the reporting scale (Beaton & Allen, 1992; Hambleton, 1998, 2002; Hambleton & Meara, 2000). Other work prioritized audience-specific differentiation (Forte Fast & Tucker, 2001; Jaeger, 2003; Levine et al., 1998; Simmons & Mwalimu, 2000), displays for state results and other subgroups (Hambleton & Slater, 1997), market-basket reporting (Mislevy, 1998; National Research Council, 2001), and digital reporting strategies (Zenisky & Hambleton, 2007; Zenisky et al., 2009). All of these efforts have contributed in significant ways to the present understanding of report development processes and report evaluation procedures.

What?

> Test scores are elusive. Even the popular percent score scale that many persons think they understand cannot be understood unless (1) the domain of content to which the percent scores are referenced is clear and (2) the method used for selecting assessment items is known. (Hambleton, 2002, p. 194)

The *what* of reporting, naturally, is about the contents of reports. Fundamentally, this is a question about data and what data elements are being communicated. It is a broad question, not only incorporating the specific results reported but also encompassing the presentation and interpretation supports (text and visual design) that play a critical role in shaping how results are communicated via reports.

As an activity that occurs part and parcel in the progression of test development, reporting encompasses the process by which data about human knowledge and skill (as measured by an instrument of some kind) are conveyed to intended users (Zenisky, 2015). In this way, the answer to the *what* of reporting is not simply a matter of scale scores, performance levels, and average scores for reference groups, but rather can perhaps be better thought of as a reflection on the key components of a report, in the sequence of the data story, referenced earlier.

In any data story, context is important, and a first way to address *what* in reporting involves the larger backdrop of the assessment context and what information broadly matters given the assessment being reported. For the purpose of results reporting, there are several key settings for reporting that are directly relevant, as previously presented in Figure 13.4: educational testing (summative and formative), workplace testing and credentialing, and psychological assessment. Each of these types of assessments accomplishes different measurement goals, and accordingly, the reports for each vary significantly in terms of their content because the underlying data and use for each differs from one another, in important and meaningful ways.

Reporting is not just about scores and data, but rather begins conceptually with information and communication. Whether the aim is content knowledge at an endpoint or at places along an educational path, demonstration of mastery to enter a profession, or an indicator of status relative to a psychological construct, these conceptualizations help to focus the reporting activity and ultimately the report product that is composed of scores and other results.

Some of these types of reports serve primarily to report an informational purpose, while others are developed with the intent to guide specific actions, and importantly, this often connects back to the purpose of an assessment. For many summative assessments in the educational context, the main reporting aim is status, and that is therefore reflected in the statement of purpose, though there may be a secondary nod to next steps in the statement of report purpose and in the report itself. In certification and licensure, the assessment purpose is to determine qualification to hold a specific credential, so reporting there prioritizes that information, though an exception here can be for failing candidates. At the discretion of a testing agency, candidates who do not pass may receive some guidance about relative strengths and weaknesses, from which the candidates can infer what might be valuable to focus on during future test preparation efforts. Formative assessment, as well as diagnostic assessment, both function at something of a crossroads in this regard. The typical aim of formative and diagnostic assessments is to inform instruction and identify areas of need, so while the reports contain information and the user of the report may or may not act on that information, the intent of the report is to provide actionable next steps. In this respect, the extent to which the aim of a report is fulfilled is up to the user.

To that end, briefly, a typical report begins with introductory content of some kind, to set the stage for the data being communicated. The nature of this introductory content may vary considerably, from individual reports to group reports, and depending on the audience. It may be as simple as a title that is descriptive of the results and some method of documenting whose results are contained within the report (as in the name of a specific individual or the parameters for inclusion in a group report, relative to geography, demographics, or other considerations). Some individual reports will go further in this regard, however, with text that describes the purpose of the assessment as well as the purpose of the report.

Another dimension of introductory content may include a statement to the effect of what is assessed, aligned to the purpose of the assessment. This should offer some broad contextual information as to the nature of test content and what is assessed. As reported by Zenisky and Hambleton (2015), a noninclusive list of examples of the descriptive elements of reports may include information such as:

- test name and/or test logo
- test date
- report title
- report purpose
- test purpose
- introductory statement from testing agency or governing body personnel
- individual reports: header space with identifying details such as name, address/school, group membership or status (individualized education plan, language, etc.)
- group reports: header space with identifying details for reported group(s) including demographic, geographical, and/or other grouping variable specifications
- details for external links to additional resources, such as curriculum materials and interpretive guides
- information about the location of frequently asked questions documents or other resources for score inquiries
- guidance on test score use/links to interpretive guidance
- glossaries of terms

The next section of most score reports is the high-level, overall, or primary results. On an individual results report, most agencies provide the overall scale score and—where appropriate—the performance level associated with that score. The formatting of these results can vary considerably, where different agencies may use different visual strategies to communicate the overall results data for one or more content areas, such as typography and font size, but also tables, bar graphs, and/or line graphs. For group reports, the overall data may be presented in a summary form, using a table or a graphic structure to communicate specific points about the data. The high-level results, whether on an individual or a group report, in the general case typically provide the results that align best to the primary purpose of the assessment. Below is a nonexhaustive list of types of scores that can serve as the basis of the scores reported (Zenisky & Hambleton, 2015). Some of these types of scores are more common in some testing applications than others:

- raw scores
- scale scores
- percentage correct
- percentiles
- stanines

- grade equivalent
- *T* scores
- performance classifications (e.g., *Advanced*, *Proficient*, *Basic*, and *Below Basic*)
- subtest or subdomain scores
- item-level scores
- student score growth

As part of the overall results, many individual reports will include additional contextual data in the form of comparisons to relevant reference groups. For educational tests, this typically means that the results for the individual student of interest may be presented alongside average scale score results for similar test takers in relevant geographic units (class, grade, city, state, country) or demographic groupings. Though many tests in the early 21st century are criterion-referenced and the primary interpretation advanced by agencies is performance relative to standards or benchmarks, these pseudo-norm-referenced comparisons are common, and often even expected. Many report users desire information about how the subject of an individual report is doing relative to others, even if little or no data about the distribution of performance are provided. It should be noted that in recent years, educational assessments have moved away from norm-referenced to criterion-referenced tests, but some form of normative data as described here is still present on many results reports. A number of examples of individual student reports can be found on the current websites for U.S. state education departments, as well as within Goodman and Hambleton's (2004) paper.

The next section(s) of many reports offer finer grained results. It is in this section of reports that, when these data are provided, report users begin to understand their performance on a test and why they achieved the score or result that they did. These finer grained results are what users look to in order to improve. The typical format of these results is the presentation of subscores, but this is a challenging topic for psychometrics because of the issues of dimensionality and the reliability of subscores and similar subdomain-level results. Most summative assessments are simply not built to offer high levels of reliability at the subdomain level, because they are intended to offer reliable results at the overall level given their articulated testing purpose and do not contain sufficient items to attain psychometric reliability for subsets of items. This is an active area of psychometric research (see Sinharay et al., 2019, for a thorough review of the considerations and issues), but a few recommendations from Sinharay et al. (2019) include the following:

- When possible, approaches such as evidence-centered design or assessment engineering are preferable where subscores are to be reported to address issues around the relative distinctiveness of the subscores since they conceptually support differentiation, because content blueprints alone may not avoid high correlations among subscores (Sinharay et al., 2007);
- Reported subscores should exhibit psychometric properties of adequate, reliability, validity, and distinctiveness;

- Agencies may wish to consider using weighted averages or augmented subscores for diagnostic purposes, rather than subscores to address the quality considerations of reliability, validity, and distinctiveness;
- When sufficient items are present for subscore reporting, subscores can be reported on the scale score metric for a test, and subscores can be equated so that interpretation of subscores can be interpreted and used consistently over forms;
- The quality of subscores aggregate over test takers can be useful but should be evaluated for indicators of quality (reliability, validity, and distinctiveness); and,
- The expectations around subscore quality for large-scale summative assessments can and should be applied for formative assessments.

It should be noted that many reports do include some form of subscore or subdomain results to provide a minimum of guidance for test takers, particularly those who obtained lower scores and who may seek information on how to improve. An additional set of recommendations around making choices as to which subscores to use and when from Sinharay et al. (2019) are as follows:

- Reporting of observed subscores may be reasonable if the subscore is reliable and dimensionally distinct.
- If evidence for subscores does not support their reliability and distinctiveness, then no amount of statistical adjustment can address fundamental limitations of the scores, and thus reporting of subscores cannot be justified.
- If subscores are moderately reliable and moderately correlated, then statistical adjustment may help.

Whereas simple subscores are easy to compute and are easily understood, Sinharay et al. (2019) do provide a comprehensive list of techniques to compute and adjust subscores, as referenced above, which can be drawn on to address the unreliability and lack of distinctiveness problems. These include augmented subscores and weighted averages, Yen's Objective Performance Index (Yen, 1987), the use of cognitive diagnostic models, multidimensional item response theory (IRT) models, and scale anchoring (Beaton & Allen, 1992). It is also certain that as data mining and data analytics work advances, so too may new methods for subscore reporting emerge.

Clearly, this topic, as an issue of psychometrics and operational practice, is not a settled matter. Large-scale assessments rarely, if ever, meet the numbers of items needed to obtain high values of traditional reliability in subscore reporting. For example, Sinharay et al. (2019) suggested that a threshold of 0.8 for reliability (based on Nunnally, 1978) is reasonable, and to get there might require 20 items. This raises philosophical issues of (a) what constitutes quality in subscores (when viewed through the lenses of reliability, validity, and distinctiveness and how those are defined) and (b) potential for harm (what are the consequences of reporting subscores that do not meet such thresholds/definitions of quality?).

Detouring briefly to group-level reports, formats typically include list-style reports of individual performance or reports that highlight summative statistics that describe

performance in aggregate. It is important to note that as a matter of good practices, the *Standards* (AERA et al., 2014) do suggest that contextual information is critical for reporting on groups, especially when reporting on group differences. For example, Standard 12.17 suggests that reports should include relevant contextual information to support meaningful and appropriate interpretations. Noted here as well is a hybrid approach to group reporting when such results are accessed online. Often such reports are targeted to instructional users and mix levels of aggregation. For example, a class-level report for a teacher may have class-level aggregated results but also display a roster-type view. In that roster view, a teacher might want to carry out certain functions, such as sorting a student list according to performance on the assessment, or click on individual student names and see more detailed individual results. The final section of a report often includes some kind of concluding information and/or guidance for next steps. Once a recipient of a report has that report in hand (physically or viewing onscreen), many reports conclude with a brief section on what might be next for the user, particularly in the context of the individual reporting. Such report sections might offer some kind of summary about performance and links to various types of resources available locally or online. The idea is that once the report document has provided some information about the performance of a test taker, the stakeholder may, for example, want to find out more about the test and use the overall and subdomain results to collaborate and develop a plan for next steps and potential improvement. Similarly, this concluding section of reports might also have links to other external documents, such as frequently asked questions where stakeholders can access additional information. Whatever such documentation is linked to from a report, whether it be lesson plans, other instructional materials, or guidance of an administrative or informational nature, it should be noted that it is incumbent on the report agency that is promoting those materials—by means of such links—to articulate the purpose of such materials; establish standards for their quality, accessibility, and usability; and provide users with logical connections to align results with provided materials, where appropriate.

Concentric Circles of Results

One visual way to consider the *who* and *what* of reports, spanning the range of individual and group reporting efforts previously discussed, is shown in Figure 13.8. This idea of the concentric circles of interest in assessment results emerges from the audience-specific considerations and research road map illustrated in Figure 13.3, and as such, Figure 13.8 further provides a general framework for thinking about the progression of audience and results. It is important to understand that not all circles of interest are necessarily applicable for all tests. Beginning at the center of Figure 13.8, the results reported are at the most local and personal level—the individual test taker. A report at that level is crafted with the test taker as the reporting unit of interest, developed for the test taker (and perhaps their family, depending on testing context), and the report contents are scores and other forms of results, often linked to targeted or specific next steps of some kind. Moving outward, the next circle of results is intended at the level of

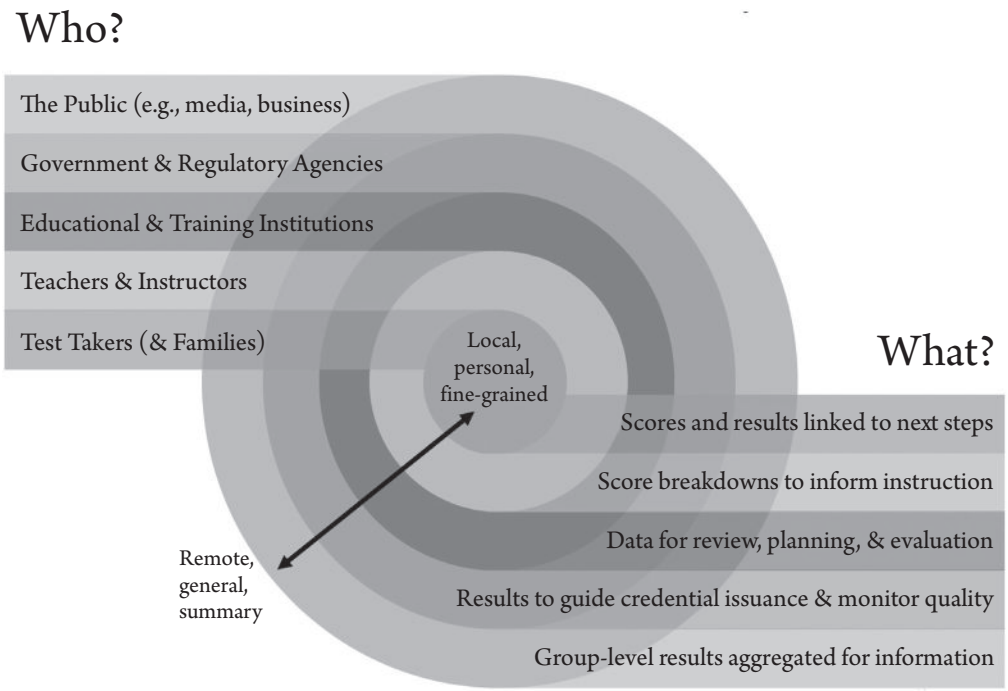


FIGURE 13.8
Concentric Circles of Interest in Assessment Results

teachers and instructors. Results at this level are still fairly local and focused on the individual, but also start to include relatively small but meaningful groupings (meaningful as defined by the user) given uses appropriate to that teacher/instructor role.

The next level of the concentric circles in Figure 13.8 is that of a progressively more general level of results, where results are communicated at the level of educational and training institutions. Here, the specific nature of individual results, and breakdowns of individual results, are typically far less consequential than a view of the results as data in aggregate to carry out goals such as review, planning, and evaluation. At this level, the performance of groups of test takers is typically analyzed and actions are identified based on the results.

Beyond that, moving further outward is the governmental and regulatory space for reporting, where individuals primarily matter for the (necessary and important) purpose of issuance of credentials. More of the results of interest at this level are for groups, with the assessment data fulfilling an informational/status purpose. Very little at this level is fine grained—the focus is on overall performance classifications for individuals and groups. The final, outer circle of interest is the public-facing side of assessment. This is the level of the kinds of databases and analysis tools such as the Data Explorers of TIMSS, PISA, PIRLS, and NAEP. Assessment results are completely scrubbed of personal identifiers and reporting is only at an aggregate level, with sample size suppression rules or bootstrap sampling methods used to shield the identification of individuals.

The overarching idea of these concentric circles is to encourage researchers and report developers to think about reporting efforts as a progression in communication. Reporting, for many tests, begins with the test taker and extends outward to other stakeholders. The nature of the results of interest necessarily change as stakeholders change, because the *use* of the results is different at each level.

Where?

When applied to results reporting, the question of *where* can be applied to format, as in, "Where are the results?" For many test takers, their results are detailed in the individual report, which is crafted to provide individual-level results in a highly structured way. Typically, such reports are static in the sense that whether disseminated on paper or onscreen via email or a secure portal, they are delivered as tightly arranged documents, with a specific sequence of results and information for every test taker, following the issuing agency's template for the data story of interest. Rarely are individual reports for large-scale summative assessments marked by any measure of user choice in content or appearance; that is increasingly common for formative assessments.

This stands in contrast to group-level results: Though historically group reporting has filled volumes of expansive reports (e.g., the "Publications and Products" library of NAEP, https://nces.ed.gov/pubsearch/getpubcats.asp?sid=031), in recent years online tools for group reporting have proliferated (e.g., the NAEP Data Explorer, https://www.nationsreportcard.gov/ndecore/). In the former case, prospective users of the information obtain PDFs of static, large-scale results written by issuing agencies for reading and reviewing, while in the latter case, the users are in the position of exploring the data within the parameters of the publicly available tools, to answer their own questions about the data by, for example, selecting who is included, what results are of interest, and how those results are displayed (Zenisky & Hambleton, 2012b).

Interactive group-level reporting tools such as the NAEP Data Explorer and the International Data Explorer (http://nces.ed.gov/surveys/international/ide/) offer unprecedented customization, placing users in charge of pursuing their own data stories. This shift in how the reported narrative is constructed can be useful and empowering, especially to data-savvy users. That said, it can also be overwhelming and pose challenges to promoting appropriate interpretations of results. Those involved in developing online reporting tools should use tutorials and design elements to guide users and restrict access to unreasonable variable selections or analyses. Other strategies in this regard might include establishing a process for receiving and responding to inquiries, displaying data in multiple formats to support different processing preferences, and using a streamlined interface to the extent possible (Zenisky & Hambleton, 2007, 2012b).

One type of reporting tool that has been gaining traction is dashboards: highly visual interfaces that provide at-a-glance views of key information, often with some interactive features. Most dashboards currently available summarize learning analytics data for teachers (i.e., metrics of student progress and engagement, like time spent on online activities, grades, and number of outputs, such as discussion posts), but a growing

number are designed specifically for students (Schwendimann et al., 2017). The use of dashboards for reporting is variable: While some groups have been using them for many years (Brown, 2001), others have been more recently deployed in educational settings, as data mining, data visualization, and web-based tools have emerged in various disciplines to fulfill visual and functional aims in terms of data communication (Sarikaya et al., 2019).

There is growing research on ways to leverage learning analytics and the interactivity of dashboards to provide more customized results and recommendations (Verbert et al., 2013). Within the context of a formative assessment integrated with instruction such as Brightpath (https://www.brightpath.com.au/formative-writing-assessments/), users can interact with dashboards that connect information to specific next steps with real-time updating of student performance and even teacher evaluations. Another interesting application for dashboards is the use of open learner models, which are visual representations of what a student knows, as well as areas of difficulty that can be updated dynamically as more data become available (Bodily et al., 2018; Kay, 1997; Zapata-Rivera, 2021). For instance, a student receiving a static Grade 7 mathematics report may see that "understanding statistical variability" is an area for improvement and make a mental note of this result. In contrast, a student accessing a dashboard with an open learner model showing the same result could interact with the model by opting to answer additional questions to support their learning and, in turn, update their status on that content area (this interactivity is what makes the model open).

As a specific format within the realm of results reporting, with additional research dashboards can become a popular alternative or complement to static reports as consumer-facing, web-based applications with powerful data aggregation features that offer real-time indicators. Consider, for instance, the dashboards from exercise- and sleep-tracking apps accessed daily by millions of smartphone users, to obtain at-a-glance reporting of quantitative measures of health up to the minute of checking (e.g., steps taken, hours slept, heart rate). A key feature of dashboards, then, is this idea of real-time reporting, and in that respect data dashboards are perhaps best suited for use in results reporting contexts aligned with formative or diagnostic initiatives that update regularly and relatively frequently.

Since dashboards for assessment results offer a middle ground between static reports and highly customizable reporting tools and serve to meet a specific real-time reporting goal, it is important to determine which elements should be interactive and which should not, and this should align with the goals of the assessment program. Developers should also consider general guidelines from the dashboard literature, such as using customization to promote a sense of empowerment and agency (e.g., allowing students to set their own performance targets and to decide what normative information they want to see, if any), designing each element to inform a particular decision or set of decisions (e.g., ensuring each dashboard section has a purpose, such as deciding which content areas warrant additional review), and using multiple visualization approaches

(e.g., color, spatial position, and motion) to support rapid perception (Bennett & Folley, 2019; Few, 2006; Park & Jo, 2019). Last, those involved in creating dashboards may adapt recommendations and approaches from the results reporting literature, combining insights from broader dashboard research with specific guidelines for reporting assessment results (Corrin et al., 2019; Kannan & Zapata-Rivera, 2022; O'Donnell et al., 2021).

When?

In the terms of the data story, the *when* of results reporting is a fairly straightforward matter of now or later. Now, in this sense, is immediately at the conclusion of test administration. With the widespread use of computerized administration, testing agencies can provide test takers with their results before they leave their seat, displayed onscreen, and/or with a report of scores emailed to them, provided that the necessary scoring procedures and quality control checks are done in those brief moments. The "later" of score reporting can range from a few days to a few months because agencies may need time to score and verify performance. In large-scale summative educational testing in the United States, for example, it is not uncommon for results to be made available to districts after 2 to 3 months and perhaps take an additional 2 months to be distributed to the students themselves.

The dimension of *when* in reporting varies considerably depending on testing context. In credentialing, the results for a clear majority of tests are reported relatively quickly, where candidates receive their scores immediately or very shortly after the test session, and results are likewise seamlessly transmitted to certification and licensure bodies. In education, formative tests tend to report scores quickly, but as described, summative test reporting tends to take a great deal longer. The time-consuming nature of scoring (and, hence, delays in reporting) can be attributed to a few causes, such as the need to human score certain item types, to perform psychometric analyses, and to do other quality checks on the items and the scores, but such delays in reporting results are also a relative impediment to the use of scores (Brown et al., 2019). Per the *Standards* (Standard 8.8), the professional guidance around the *when* of testing indicates that results should be provided in a "timely" manner, but what constitutes timely may be a matter of debate and can be viewed in light of official test purposes. In some testing contexts, reports of results are expected and used quickly (such as entry into a profession), and in other contexts, longer intervals between testing and reporting may be more tolerated.

Why?

Arguably, the most interesting question of the data story metaphor as applied to results reporting is *why*. In the *Standards* (AERA et al., 2014), Standard 5.1 speaks to the fundamental requirement that test users should be provided with clear expectations of the characteristics, meaning, and intended interpretations of scores, but it falls to two other standards (8.8 and 9.16) to establish the underlying expectation that scores themselves are reported, particularly with respect to test takers.

> Standard 8.8: When test scores are used to make decisions about a test taker or to make recommendations to a test taker or a third party, the test taker should have timely access to a copy of any report of test scores and test interpretation, unless that right has been waived explicitly in the test taker's informed consent document or implicitly through the application procedure in education, credentialing, or employment testing or is prohibited by law of court order. (p. 136)
>
> Standard 9.16: Unless circumstances clearly require that test results be withheld, a test user is obligated to provide a timely report of the results to a test taker and others entitled to receive this information. (p. 146)

These two standards set a baseline for reporting to test takers, conditional on use (per Standard 8.8) and circumstances (Standard 9.16). If scores have implications for an individual, then individuals must be provided with their results report in a timely manner. Taking Standards 8.8 and 9.16 along with Standard 5.1 (described previously), it is clear that when scores matter, they must be provided along with guidance around interpretation and use.

However, the process of reporting results happens in line with test purposes (e.g., educational summative, educational formative, credentialing, admissions, psychological). This is an important point to maintain in the conceptualization and development of results reports for various intended audiences. Often, the *use* of scores falls to entities other than the test taker, and accordingly the nature of the report sent to those external entities hinges on what they plan to do with the data. In many—not all, but many—settings, reporting to the actual test takers simply involves a report that is informational in nature, and use is left to others. To this end, in those cases, the *why* of reporting depends on the intended user (the *who*). In effect, reporting to test takers in many cases fulfills the question of *why* with an informational aim, rather than an actionable use by the test taker. Then, for other, external audiences, the *why* of reporting is generally more aligned to the fundamental purpose of the test (to demonstrate competence in a profession, to rank order individuals to inform admissions decisions, to determine mastery of content for granting a diploma, etc.).

The main exception to this, of course, is in education. The reality of reporting in the 21st century is that most users of test data in educational settings have expectations for reporting that extend deeper into the data, beyond high-level performance characterizations and toward formative or instructional purposes, which are not purely informational in nature, nor are they often the validated purpose(s) of the test (see Brookhart & DePascale, this volume)—they are *actionable*. The *why* of reporting to these users is different yet again, in that their needs and interests in the data are centered on instructional planning and support at micro (individual) or macro (group) levels (Brown et al., 2023).

How?

Discussion of the methods and practices of score reporting would not likely be complete without some attention paid to the *how* of reporting, in this case meaning the

psychometric methods that are used to produce the scores and other results provided in reporting documents. Such methods can span from relatively simple descriptive values to scores that are obtained through complex statistical modeling approaches. In some ways, this is closely entwined with the *what* of reporting, in that it speaks to how scores and other results may be computed from test data. Of course, the *how* of obtaining scores depends on the source data and the reporting aim for those particular data. A few strategies for producing scores and other results are provided in the following pages, relative to reporting aims, reflecting a broad continuum of computational and conceptual strategies.

Raw scores are perhaps the simplest of approaches: They represent the number of points obtained out of the number of points possible, with no weighting or transformation done. Raw scores are often used in informal or classroom settings, but are not typically used to report overall test results on large-scale standardized tests for various reasons, including to ensure comparability of scores across forms where minor difficulty differences are detected or to minimize preconceived notions of the reporting scale (Hambleton & Zenisky, 2003). Some large-scale assessments have reported subdomain results using raw scores—or a simple derivative thereof, such as percentage correct—per Goodman and Hambleton (2004). Scale scores are a step up from raw scores in that some type of transformation is applied for reporting purposes, such as linear or nonlinear. Raw scores, and scale scores, can easily be averaged over test takers for local or more global indicators of performance, and in this way scores for individuals can be interpreted in norm-referenced ways against sample or population averages. Similarly, cut scores can be computed and applied to raw or transformed scores for grouping test takers into achievement levels (such as pass/fail or gradations of proficiency, such as those used in educational assessments and represented earlier in Figure 13.4).

IRT likewise offers much in the way of reporting. IRT can be used not only for the purpose of estimation, but also for reporting, because item difficulty and test-taker proficiency are on the same scale, which can be mobilized to add meaning to the reporting scale (Hambleton & Zenisky, 2017). Some especially powerful examples of this include the item mapping approach used by NAEP (see U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2015), the score reports used on the Massachusetts Adult Proficiency Test (Zenisky et al., 2018), and the analogous Rasch-based Wright map (Measurement Research Associates, 2010). Item mapping is conceptually connected to the work by Beaton and Allen (1992) on scale anchoring, where specific points on the reporting scale, the anchor points, are illustrated by test content.

Because testing technology has been and is ever evolving, the data that can be used for the *how* of reporting are likewise expanding. Crowdsourced learning (Milligan & Griffin, 2016) serves as but one example of the use of logstream data for scale development and, ultimately, reporting. Indeed, efforts to report data based on not just scores

but also process data offer clear innovations for the *how* of testing (e.g., De Boeck & Scalise, 2019; Ercikan & Pellegrino, 2017; Xiao et al., 2022), as does advancing work in the field of learning analytics (e.g., Du et al., 2021; Oliva-Córdova et al., 2021; Siemens & Baker, 2012). As agencies consider scores that derive from other data sources, the meaning of scores and other results will continue to shift and reporting will accordingly need to change as well, to reflect results that are increasingly multidisciplinary and indicative of a different orientation in conceptualizing and representing performance.

Report Development Process

> Quality score reporting does not happen by happy accident. It requires commitment from testing agencies to a process that is closely integrated into the larger schedule of test development activities, to define the necessary connections between the choices made in test development and the validity of the test score inferences to be communicated. (Zenisky & Hambleton, 2015, p. 601)

To this point, much of this chapter has focused on conceptual perspectives on reporting and to establish reporting as an essential activity in the tradition of the IUA of validity. With this grounding, the focus now turns to report development in earnest, and specifically models for crafting reports using thoughtful and collaborative processes. Models developed by Zapata-Rivera (2011) and Hambleton and Zenisky (2013) can be used to guide the report development process. Zapata-Rivera's (2011) model includes four main steps: "(a) gathering assessment information needs from stakeholders; (b) reconciling these needs with the available assessment information, (c) designing various score report prototypes, and (d) evaluating these score report prototypes internally and externally" (p. 37). We focus on the Hambleton and Zenisky (2013) model for the remainder of this section, but also highlight recommendations from Zapata-Rivera that apply to different stages of the development process.

Zenisky and Hambleton (2012a) introduced an initial version of the Hambleton and Zenisky model and it was described in more detail in later publications (Hambleton & Zenisky, 2013; Zenisky & Hambleton, 2015). The model was informed by knowledge of the reporting literature, best practices in test development, and direct experience with report design. Further, it stemmed from the idea that reporting efforts, like other processes including item analysis and standard setting, should follow a general cycle of development. The model was designed to be flexible and to apply to various testing contexts, so it emphasizes process considerations rather than rigid content and design requirements. As such, the model may be adapted to guide the development of reporting tools (e.g., dashboards for assessment results), as demonstrated by O'Donnell et al. (2021). Readers interested in developing such tools are encouraged to consider the Hambleton and Zenisky model, replacing "report" with "reporting tool," as well as literature from the learning analytics community on evaluating the usefulness, usability, and effectiveness of dashboards (see Corrin et al., 2019, for a review).

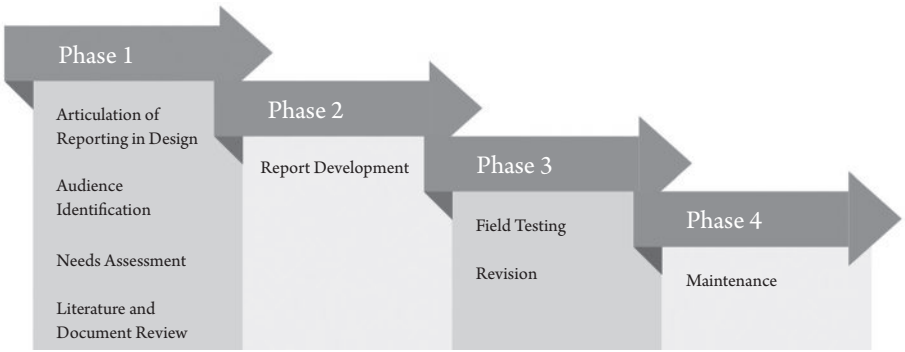


FIGURE 13.9
The Hambleton and Zenisky Model for Score Report Development

Note. Adapted from "Reporting Test Scores in More Meaningful Ways: A Research-Based Approach to Score Report Design," 2013, by R. K. Hambleton & A. L. Zenisky, in K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C., Hansen, N. R. Kuncel, S. P. Reise, and M. C. Rodriguez, *APA Handbook of Testing and Assessment in Psychology: Vol 3. Testing and Assessment in School Psychology and Education* (pp. 479–494). American Psychological Association. https://doi.org/10.1037/14049-023

At a high level, the Hambleton and Zenisky model (Figure 13.9) begins with four tasks: articulating how the test, results, and reports will jointly support intended inferences (1a), identifying all groups of intended report users (1b), completing a needs assessment for each of those groups (1c), and conducting a literature and document review (1d). The next steps involve using the information obtained previously to develop prototypes (2), completing as many iterations of field testing and revisions as needed for reports to be ready for operational use (3), and implementing a maintenance plan after the reports or reporting tools have been released to the public (4).

We describe each element of the model in more detail in the following paragraphs. But, before delving into the *how* of report development, it is important to consider *who* should be involved in the process. A variety of talents and perspectives are needed to tackle reporting challenges and foster innovation. Reporting is an interdisciplinary activity, and so teams of report developers should also be interdisciplinary. As described by Slater et al. (2019), individuals on report development teams may include graphic designers, user experience practitioners, cognitive science researchers, psychometricians, assessment developers, information technology staff, and accessibility experts. Teams may include developers who serve multiple roles, and consultants may be brought in if needed. Collaboration is key in each step of the development process, as described through the lens of the Hambleton and Zenisky model.

Phase 1: Laying the Groundwork

The first element of the Hambleton and Zenisky model, *articulation of reporting in design* (1a), requires deliberate consideration of the alignment among the purpose(s) of the test, test design features, and the aims of the report(s) under development. It cannot be overstated that reporting goals should be considered early in the test development process to ensure that decisions about content, format, and test length are in agreement with the kinds of information an assessment should provide. If there are distinct claims

about reports and score interpretations that are specified at the beginning of the test development process, then reports that are intended to support the IUA should likewise be framed out at that same time to ensure that test design choices actually gather and provide the data to support the eventual reporting activities and report contents, in effect a backward mapping of reports to the test development process. At this initial stage, it may be helpful to begin preparing a "prospective report"—a simple mock-up of what the final report might include that can be refined as more information becomes available and used to inform test development discussions (Zapata-Rivera et al., 2012).

The next element, *audience identification* (1b), involves naming the user groups that will need information from the report(s) to draw conclusions or take specific actions. In the K–12 context, the target users for a test taker–level report may be students and parents, while teachers and administrators may be the target users for a group-level report. In the credentialing context, those candidates scoring above the minimum passing standards and those below the minimum passing standard may perhaps be considered distinct user groups, because their reporting interests may be quite different depending on what side of the standard they are on. Jaeger's (2003) work illustrated the multiple user groups that may rely on reports from an assessment program—up to nine groups in the case of U.S.-based NAEP. It must be understood as well that in the process of identifying user groups, such groups are composed of individuals, who bring varying interests, background knowledge, and needs to assessment data communication. In this step, audiences must be differentiated, but the variation within each user group must also be acknowledged.

Thus, after identifying the target audience, a *needs assessment* (1c) should be conducted to understand each user group's information needs, as well as characteristics that might influence how members of the group approach assessment results. From an audience analysis perspective, Zapata-Rivera and Katz (2014) recommended focusing on the following areas:

- Needs: Users' purpose for reviewing a report or reporting tool, exemplified by questions such as, "What inferences about or decisions concerning test takers does the user want to make?" (p. 447)
- Knowledge: What users already know and what information they need to fully comprehend the report, exemplified by questions such as, "What knowledge gaps might interfere with correct interpretation of the score report?" (p. 447)
- Attitudes: The feelings or biases that might influence users' interpretations of the information reported, exemplified by questions such as, "What do users think about assessment generally?" and "What are their expectations about the test takers?" (p. 447)

The needs assessment should be viewed as a broad-based inquiry, extending well beyond a simple investigation of user groups' familiarity with statistical terms (Zapata-Rivera & Katz, 2014; Zenisky & Hambleton, 2015). It must take place early in the test development process, to inform test development decisions and be informed by user reporting

needs and interests. Zenisky and Hambleton (2015) also recommended gathering general feedback on how users would like to access the report and what data elements they would like to see, which can then be reconciled with what the assessment program is able to support. Report developers may collect information for this step directly through focus groups or surveys or indirectly by reflecting on their own experience with the target audience. If an early prospective report was created, it can be shared with stakeholders, if appropriate, and refined based on key feedback from this stage.

The last element in Phase 1 of the Hambleton and Zenisky model is a *literature and document review* (1d). Report developers should be familiar with general principles from the literature (e.g., Hattie, 2010; Slater et al., 2019; Zenisky & Hambleton, 2012a), as well as findings specific to report components under consideration, such as visual displays (e.g., Hegarty, 2019; Ryan, 2006; Wainer, 1997b), language (Roduta Roberts et al., 2018), or representations of measurement error (e.g., Kannan et al., 2018; Zapata-Rivera et al., 2019; Zwick et al., 2014). It is also helpful to note which methods have been employed to gather feedback on reports, paying special attention to those involving similar assessments and/or audiences.

A number of organizations such as state departments of education to credentialing agencies regularly post sample reports and interpretive guides online (Knupp & Ansley, 2008). To learn from current design approaches, it is useful to examine those samples (or reporting tools, if applicable), keeping track of content and design choices of interest based on the assessment data available and audience characteristics. There is no combination of content and design elements that guarantees universal success, but reviewing a series of complete reports and interpretive guides helps the report development team in the process of beginning to imagine the possibilities (Goodman & Hambleton, 2004).

Indeed, interpretive guides play a pivotal role for report users seeking additional guidance. Reports should be self-contained and concise, including all essential information to support accurate inferences. However, paper or online interpretive guides should still be available for users who may not be as familiar with the context of the assessment or who may be interested in learning more. Goodman and Hambleton (2004) identified several interpretive guide components for consideration, including answers to common questions about the assessment, more information about its content and purpose, suggestions to improve performance, and guidance on where to find additional resources.

Phase 2: Report Development

In Phase 2, *report development*, the information gathered earlier in the process should be used to create a set of prototypes that are "aligned with the test's goal(s), audience-specific and also rooted in best practices" (Zenisky & Hambleton, 2015, p. 593). Crafting multiple prototypes is helpful so that members of the target audience can react to a range of content and design options. Although it is not impossible, it seems unlikely that a group of report developers would get all elements right on the first and only prototype developed.

The prototypes may have text and design elements in common, and it is up to the report developers to decide how detailed the initial set should be. To the extent possible, the prototypes should reflect the "look and feel" of a complete report or reporting tool and be informed by data gathered in Phase 1. Prototypes missing key features may lead to inaccurate results during field testing, such as members of the target audience indicating that a design is not cluttered only because some or all text is missing, or having trouble understanding the flow of information because a major graph or other element has not yet been added. If multiple reports are needed, a list showing the audience and core data elements for each can assist with tracking development progress.

Phase 3: Field Testing

> Always field-test graphs, figures, and tables on focus groups representing the intended audiences; many important things can be learned from field-testing report forms such as features of reports which may be confusing to readers. (Hambleton & Slater, 1997, p. 18)

Phase 3 involves an iterative cycle of *field testing* and *revisions* until the report is ready for implementation. An internal review by colleagues from different backgrounds or with sufficient knowledge of the target audience may be conducted prior to external review. Table 13.1 offers a comprehensive list of evaluation questions that may be used at this stage. The questions are grounded in best practices from the literature and are meant to promote critical reflection. Following internal review, recommendations that lead to minor changes can be readily implemented (e.g., wording suggestions to improve

Table 13.1. Updated Evaluation Form for Reviewing Reports

Report Element	Report Review Questions
I. Overall	A. What are the key intended interpretations of the report? B. How does the report reflect the interests and informational needs of key stakeholders? C. In what specific ways does the report present information that aligns to the purpose(s) of the assessment? D. What evidence has been gathered to support the validity of the report for intended user groups?
II. Content—report introduction and description	A. Does the report have a title clearly and descriptively identifying whether it is for an individual or a group? B. For group reports, how are the parameters for inclusion in the group defined? C. How are details provided about the content of the test(s) being reported, if it is not common knowledge for the audience? D. How is the purpose(s) of the test expressed, given the intended audience? E. If present, in what way does the introductory statement set a positive tone for the report?

Report Element	Report Review Questions
III. Content—scores and performance levels	A. How many different reporting scales are used on the report? Is the meaning and range of each score scale communicated, and how? B. How are the performance categories or psychological states being used (e.g., *Basic*, *Proficient*, *Advanced*) reflective of intended inferences, not stigmatizing, and described sufficiently for the intended audience? C. If it is not obvious, how is information provided to guide interpretation and use for each score and classification? D. What strategies are used to guide users away from known misinterpretations and misuses? E. How is the topic of score imprecision handled, to promote interpretation, for each result reported (i.e., overall and subdomains)? F. Have "probabilities" or "conditional probabilities" been used? If they are used, is the explanation clear? G. Is there sufficient information for the reader, without being overwhelming?
IV. Content—other performance indicators	A. Is there any linking of test results to possible follow-up activities? For example, with educational tests, are the results linked to possible instructional follow-up? B. If present, are relevant reference group comparisons reported with information on appropriate interpretations? C. If norms are provided, what steps are taken to describe the reference group in sufficient detail? D. If present, how are results of performance on individual test questions reported to facilitate use and understanding? E. If present, how are reports of scores from other recent and relevant tests explained?
V. Content—other	A. How does the report provide information about where to direct questions? B. Does the report provide links to additional resources about the test, testing program, and/or understanding test-taker performance?
VI. Language	A. What steps have been taken to ensure that the report is free of statistical and other technical jargon and symbols that do not facilitate or promote understanding and interpretation? B. What steps have been taken to ensure that the text is clearly written for users? C. If footnotes are used, are they clearly written for the reader? D. If the report (or ancillary materials) is translated/adapted into other languages, how is the translation/adaptation carried out? What steps were taken to validate the translated/adapted version?
VII. Design	A. What visual and/or narrative strategies are used to highlight the information that is most important based on the purpose(s) of the assessment? B. How is the report clearly and logically divided into distinct sections to facilitate readability?

(continued)

Table 13.1 *(continued)*

Report Element	Report Review Questions
	C. What visual or narrative strategies are used to communicate the key score information? D. Is the font size in the different sections suitable for the intended audience? E. What steps have been taken to ensure that the graphics (if any) are presented clearly to the intended audience? F. Is there a mix of text, tables, and graphics to support and facilitate understanding of the report data and information? G. What evidence has been collected to suggest that the report looks friendly and attractive to users? H. What are the steps taken to ensure that the report has a modern "feel" to it, with effective use of color and density (a good ratio between content and white space)? I. What steps have been taken to ensure that the report is free of irrelevant material and/or material that may not be necessary to address the purposes of the report? J. What evidence is gathered to suggest that the "flow" for reading the report is clear to the intended audience? K. How does the report align in layout and design to related materials published by the testing program?
VIII. Interpretive guides and ancillary materials	A. Is there an interpretive guide prepared, and if so, what steps have been taken to ensure that it is informative and clearly written? Has it been field tested? Are multiple language versions available to meet the needs of intended readers? B. If there is an interpretive guide, is there an explanation of both acceptable and unacceptable interpretations of the test results?

Note. Adapted from "Reporting Test Scores in More Meaningful Ways: A Research-Based Approach to Score Report Design," 2013, by R. K. Hambleton & A. L. Zenisky, in K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C., Hansen, N. R. Kuncel, S. P. Reise, and M. C. Rodriguez, *APA Handbook of Testing and Assessment in Psychology: Vol 3. Testing and Assessment in School Psychology and Education* (pp. 479–494). American Psychological Association. https://doi.org/10.1037/14049-023

clarity, small formatting changes), but more complex requests should be considered with caution, and it may be best to wait for external review to confirm the request is aligned with report users' wants and needs.

During external review, it is imperative that report developers choose field-testing approaches that provide information not only on preferences, but also on comprehension. The report elements that users like and the elements they understand are not always the same (e.g., Wainer et al., 1999). Different approaches yield different information, and accordingly, it may be advantageous to use more than one method to obtain data about reports using multiple approaches.

Surveys may be an efficient way to gather data on both areas and reach a high number of participants. Typically, surveys used for field testing include images of the prototypes and a mix of rating scale and open-ended questions. Participants may be more willing to provide candid answers and may enjoy being able to explore

materials and answer questions at their own pace. However, depending on their level of motivation and interest, engagement with the questions may be lower than for methods with active facilitation. Last, report developers must decide whether to make the survey anonymous or request contact information to allow for follow-up questions.

Another option is to conduct focus groups (e.g., Forte Fast & Tucker, 2001; Ryan, 2006; Zenisky, Delton, & Hambleton, 2006; Zenisky, Hambleton, & Smith, 2006). They offer the possibility of hearing correct interpretations or misinterpretation of reports, allow asking follow-up questions as needed, and can be useful for getting answers to prepared questions as well as exploring unplanned but important topics that arise during discussion. To encourage full participation, Zenisky and Hambleton (2015) recommended holding separate focus groups with members of different user groups and being mindful of potential power imbalances among participants. Probing comprehension in a group setting can be challenging; participants may be unwilling to venture interpretations or admit that something is difficult to understand. An alternative to asking direct comprehension questions is asking questions such as:

- Do you think other [parents/teachers/students] would understand this information?
- Is there something in particular that might be confusing to them?
- What is your interpretation of this report?
- What would you do next now that you have interpreted this report?

Such questions can lead to many interactions among focus group members that can inform the adequacy of the report and areas for improvement.

An additional source of evaluative questions in this regard can be sourced from Ryan (2006). These can be tailored to be relevant to the intended user group as well as the person doing the ratings.

- Will a [user] find this information helpful?
- How could a [user] use this information?
- Could this information be modified to be more informative or helpful?
- How can this information be best presented?
- Might there be any problems in how this information is used?

These particular questions were posed as conceptual points to help study participants in the context of rating different proposed reporting formats, but can also be viewed as stand-alone questions that could be asked of participants.

Interviews are the most time-intensive option, but can provide rich information on an individual's perceptions, and participants may feel more comfortable sharing honest feedback on prototype elements that are difficult to interpret in this individual setting. Like focus groups, interviews can be based on a set of prepared questions or follow a more conversational style. Unlike focus groups, they can incorporate techniques

such as think-aloud protocols, where users are asked to voice their thought processes and reactions as they interact with a prototype. When there is particular interest in understanding how users navigate prototypes, eye tracking and direct observation may also be used.

Regardless of methods used to gather data on reports, report developers must be acutely involved in the process of identifying participants for feedback on reports and seek out research participants who may be able to provide wide-ranging perspectives on report prototypes. It is not sufficient to name intended user groups at a general level (e.g., "parents," "teachers") while ignoring the subgroups that are present within such groups. Broadly speaking, sex, race/ethnicity, socioeconomic status, proficiency in the primary language of the report, and technology familiarity (for web-based reporting efforts) may provide initial sampling frames for recruiting participants. For example, in a reporting jurisdiction with a high proportion of Spanish-speaking families, it would be appropriate to develop a sampling plan that ensures representation across the diversity of language status in that jurisdiction. This approach not only benefits a specific report document but also can glean insight about communication and access strategies around reporting. Similarly, if a report is developed in certification and licensure contexts that offer a diagnostic component for failing/low-performing candidates, with the aim of providing actionable information, then failing/feedback from low-performing candidates would be something that would likely be useful to include on a sampling plan. The process of gathering feedback at any point in the report development cycle can be time-consuming and complicated, but careful consideration of the research participants, including clear articulation of relevant subgroups for any individual user group and outreach to engage those subgroups, cannot be overlooked as an aspect of validating reports. Just as test developers seek out diversity in test specifications panels, item writers, and standard-setting panels, so too must diversity be prioritized in participatory research around reporting.

After analyzing the results from field testing and identifying areas of consensus, report developers must decide which content and design changes to make. Some improvements may be obvious upon reviewing the feedback, while others may not be immediately feasible or take additional consideration. For example, there may be consensus that a particular layout is confusing, but no suggestions on how to improve it. Depending on the magnitude of the revisions, additional field testing may be needed to investigate whether the changes had the intended effect. This phase is intended to be iterative, when appropriate.

Phase 4: Maintenance

The *maintenance* phase begins once reports become operational. As explained in the validity section, investigating whether reports are interpreted and used as intended is critical. This phase may include periodic evaluations using some of the same methods employed for field testing in addition to continuous monitoring of inquiries and comments from

report users received via customer support channels, online forums, and other outlets. If content or design flaws are identified, "easy fixes" may be implemented right away, or potential revision ideas may be accumulated over a certain period to implement a number of changes at once. As with other phases of the model, there are no strict rules about how often to conduct maintenance research and what methods may be involved, but critically there should be a plan in place, and the plan should be aligned with the process for gathering evidence of actual interpretation and uses of the information reported.

Summary

The Hambleton and Zenisky model presented here, supported by the principles and ideas from Zapata-Rivera (2011), is an approach to report development that aligns with research and practice. The driving force behind this model is to ensure that results reports, as a public-facing end product of the test development process, meet the known and articulated needs of the various intended users, through a collaborative and iterative series of steps. Evaluation is a critical aspect of this work, and the checklist provided in Table 13.1 offers report development teams opportunities to reflect on their work and the final report to demonstrate its validity relative to the test's purpose and the communication process.

WHAT WE HAVE LEARNED

A theme throughout this chapter is that developing effective reports, whether static or interactive, requires careful consideration of the who, what, where, when, why, and how: specifically, who will use the results, what data are available, where the results will be reported, when they will be released, why results are needed, and how data will be turned into results. These factors are interdependent and, combined, create numerous possibilities. Teams developing reports or reporting tools may avail themselves of models such as the one in the previous section, but because effective reporting is so context specific, summaries of findings by report element may be difficult to find.

Gotch and Roduta Roberts (2018) conducted a review of 60 studies on individual reports published between 2005 and 2015, summarizing areas of focus, theoretical frameworks of communication, and data characteristics rather than findings regarding specific report elements. They noted that "data sets were often small or localized to a single context" (p. 46), highlighting the issue of generalizability. Although there have been no meta-analyses of findings from studies on results reporting to date, likely due to how context dependent such findings tend to be, some general guidance is available.

To this end, Figures 13.10 through 13.17 offer key recommendations from five sources: Goodman and Hambleton's (2004) review of results reports and interpretive guides; Ryan's (2006) *Handbook of Test Development* chapter on reporting practices, issues, and trends; Hattie's (2009) principles for promoting the validity of reports; Zenisky and Hambleton's (2012b) guidelines for developing online reporting resources; and Slater et al.'s (2019) chapter on designing reports for large-scale testing programs.

These recommendations are grouped on the basis of eight dimensions of reports and reporting: overall; content—report introduction and description; content—scores and performance levels; content—other performance indicators; content—other; language; design; and interpretive guidance and ancillary materials (these dimensions align to the report elements of the checklist presented in Table 13.1).

These works offer advice based on findings and practical experience from scholars who have dedicated substantial portions of their careers to improving the communication of assessment results. Although visualization ideas from Wainer are not directly represented, Wainer's extensive contributions and publications significantly informed the work of Goodman and Hambleton (2004) and Ryan (2006).

Goodman & Hambleton (2004)

"Consideration should be given to the creation of specially designed reports that cater to the particular needs of different users." (p. 219)
"Personalize the student score reports and interpretive guides." (p. 219)
"Reports should be piloted with members of the intended audience." (p. 219)

Ryan (2006)

"A score report should be related to content standards as clearly and explicity as possible." (p. 705)

Hattie (2009)

"The validity of reports is a function of the reader's correct and appropriate inferences and/or actions about the test takers' performance based on the scores from the test." (p. 3)
"Evidence is needed to demonstrate how readers are interpreting reports." (p. 4)
"A report should be timely to the decisions being made (formative, diagnostic, summative, and ascriptive)." (p. 13)
"Anchor the [report or reporting] tool in the task domain." (p. 9)

Zenisky & Hambleton (2012a)

Static: "Pay attention to the reporting interests and needs of users, as presenting results in static format means users see only what is shown to them." (p. 179)
Interactive: "Have users articulate their data analysis needs." (p. 181)
"Look into what others have done. Many states and publishers are being very creative in developing online interactive tools." (p. 181)

Slater et al. (2019)

"Design the report so that viewers can see and understand the most important information in 10 seconds or less." (p. 98)
"Make sure that the report design can accommodate unusual but possible conditions, such as very long names or very low or high scores." (p. 99)
"Follow Web Content Accessibility Guidelines (WCAG; Caldwell, Cooper, Reid, & Vanderheiden, 2008)." (p. 99)

FIGURE 13.10
Key Reporting Recommendations: Overall

Goodman & Hambleton (2004)

"Include an easy-to-read narrative summary of the student's results at the beginning of the student score report." (p. 219)
"Devices such as boxes and graphics should be used to highlight main findings." (p. 219)

Ryan (2006)

"Score reports should highlight important results in some way, (e.g., boxes, boldface type)." (p. 706)

Hattie (2009)

"A report should provide justification of the test for the specific applied purpose and for the utility of the test in the applied setting." (p. 11)

Slater et al. (2019)

"Emphasize the most important information in the score report. . . . The most important parts of the report should command the most attention." (p. 98)

FIGURE 13.11
Key Reporting Recommendations: Content—Report Introduction and Description

Goodman & Hambleton (2004)

"Include all information essential to proper interpretation of assessment results in student score reports." (p. 219)

Ryan (2006)

"[Results] should be reported in relation to performance standards. " (p. 705)
"[Results] should be reported at the finest level of detail for which reliable information can be provided." (p. 706)
"A score report should include information about precision for all scores presented." (p. 706)

Ha ie (2009)

"Maximize interpretations and minimize the use of numbers." (p. 5)

Zenisky & Hambleton (2012a)

Static: "Try out different report formats (summary, highlights, full-length reports) and data displays (text, graph, and tables) with intended audiences to ensure that materials are understood and the conclusions being drawn are appropriate. " (p. 179)
Interactive: "Start with things that are manageable, and look into developing comparatively simple web interfaces supported by databases that let users select results for a content area, unit of analysis (school, city/town, region/county, state), and perhaps limited demographic characteristics." (p. 181)

FIGURE 13.12
Key Reporting Recommendations: Content—Scores and Performance Levels

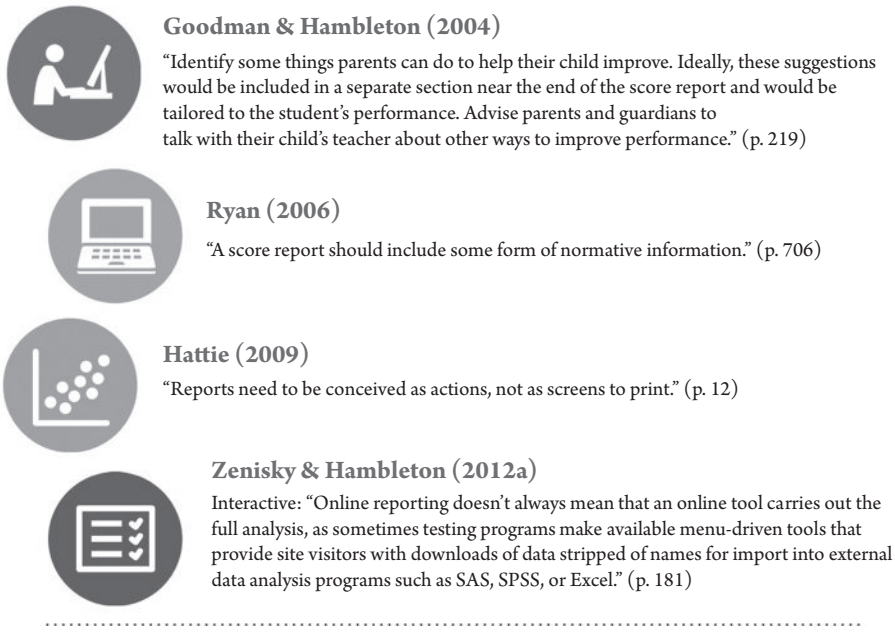


FIGURE 13.13
Key Reporting Recommendations: Content—Other Performance Indicators

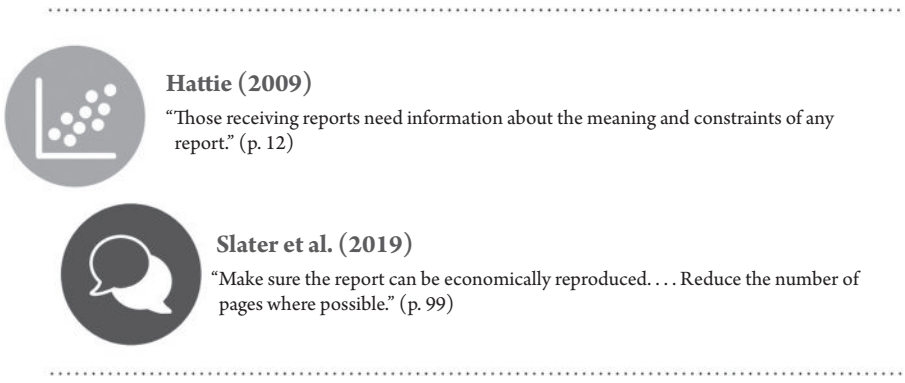


FIGURE 13.14
Key Reporting Recommendations: Content—Other

The Road Ahead: Opportunities in an Evolving Reporting Landscape

While much research and effort has been exerted to formalize reporting efforts within the broader domain of test development, the ongoing reimaginings of assessment spurred in part by technological advances within and outside assessment demonstrate that reporting is a topic that itself is not static, and work in this area is ongoing. In this section, several priority areas for future work are briefly highlighted.

Goodman & Hambleton (2004)

"Score reports should include easy-to-read text that supports and improves the interpretation of charts and tables." (p. 219)
"Small fonts, footnotes, and statistical jargon should be avoided." (p. 219)
"Key terms should be defined, preferably within a glossary." (p. 219)

Ryan (2006)

"Avoid jargon unfamiliar to the intended audience." (p. 706)
"Provide an explanation or glossary for any measurement terms used." (p. 706)
"Use text to explain graphs, charts and tables." (p. 706)

Slater et al. (2019)

"Ensure that any language in the report is appropriate to its intended audience, at the proper reading level." (p. 98)
"Avoid technical language that might be difficult for non-experts to understand." (p. 98)

FIGURE 13.15
Key Reporting Recommendations: Language

Goodman & Hambleton (2004)

"Score reports should be clear, concise, and visually attractive." (p. 219)
"Care should be taken to not try to do too much with a data display (i.e., displays should be designed to satisfy a small number of preestablished purposes)." (p. 219)
"Data should be grouped in meaningful ways." (p. 219)

Ryan (2006)

"Score reports should be clear, as simple as possible, and uncluttered." (p. 706)
"Use simple and clear graphs, charts, and tables." (p. 706)
"Score reports should use print features such as font size, style, and spacing that make it as easy as possible for the reader to understand the report." (p. 706)

Hattie (2009)

"Each report needs to have a major theme." (p. 7)
"A report should be designed to address specific questions." (p. 11)
"Readers of reports need a guarantee of safe passage… and destination recovery." (p. 4)
"The answer is never more than 7 plus or minus two." (p. 6)
"A report should minimize scrolling, be uncluttered, and maximize the 'seen' over the 'read.'" (p. 10)

Zenisky & Hambleton (2012a)

Interactive: "Different users have different capabilities for understanding quantitative data, so look into presenting data in multiple formats when possible (tables, graphs, narrative, etc.)" (p. 181)

Slater et al. (2019)

"Create a strong visual hierarchy that guides the viewer's eye appropriately through the report" (p. 98)
"Eliminate visual clutter . . . Avoid repeating elements of the report, except where repeating them makes the report clearer." (p. 98)
"Avoid using lines that add to visual complexity. Instead, use shaded areas to delineate space." (p. 98)
"Use visual embellishments such as icons only when they make it easier for users to correctly interpret the report. . . Avoid any visual element that may have a negative connotation." (p. 98)
"Colors, used meaningfully, can make the report easier to interpret. However, make sure that the report can convey all its information even if the viewer cannot accurately differentiate colors or if the report is printed or photocopied in black and white." (p. 99)

FIGURE 13.16
Key Reporting Recommendations: Design

Goodman & Hambleton (2004)

"Include detailed information about the assessment and score results in a separate interpretive guide." (p. 219)
"Include sample questions in the interpretive guides that illustrate the types of achievement represented by each performance level." (p. 220)
"Include a reproduction of student score reports in the interpretive guides to clearly explain the various elements of the reports." (p. 220)

Zenisky & Hambleton (2012a)

Informational Web Pages:
"Ask stakeholders what information about the testing program would be helpful for them to know." (p. 182)
"Review other testing programs' websites to find out the kind of programmatic documentation they make available." (p. 182)
"Post the technical manual and updates online as available." (p. 182)

FIGURE 13.17
Key Reporting Recommendations: Interpretive Guides and Ancillary Materials

Reporting in the Formative Assessment Context

Reporting results from formative assessments is a growing area of research and operational development. While the focus of most reports for summative interpretations is a total score indicating overall achievement in relation to a target construct, such a broad indicator of performance is typically much less interesting and useful in formative contexts. Formative interpretations use a range of formats and delivery methods, but their defining characteristics are being administered *during* instruction and providing feedback to inform adjustments to teaching and/or studying approaches. Informing immediate action to support student learning is central to formative assessment, creating unique challenges and opportunities for reporting.

While expediency is desirable in all testing contexts, it is essential to formative interpretations: results must be ready in time to make a difference. Rapid reporting increases the relevance of the information provided and the probability that educators will act on the results (Brown et al., 2019). Specificity is also important. Reports for formative interpretations should provide sufficient detail so that users can identify areas of strength and weakness and know where to focus to make improvements. These finer grained results have traditionally included item-level feedback and scores or achievement levels for specific content areas or attributes. More recently, the options have expanded to include process data.

Process data may include time spent on each assessment task, logs of which resources were used at different points, logs of answer changes, click streams, and keystroke logs. Turning such data into insights holds substantial promise, but perhaps most especially in formative contexts, because it can paint a richer picture of student learning. An example of this is found in Milligan & Griffin (2016), who elaborated a generalized six-step methodology for constructing measures using click-stream data. However, research on how to do this, how to report process data, and how to establish the validity of this practice is still in its infancy, and there is much yet to be learned (Provasnik,

2021; Zapata-Rivera et al., 2021). Analysis approaches drawn from the cognitive diagnosis modeling literature, such as the attribute hierarchy method, may also be helpful here (Roberts & Gierl, 2010).

With or without process data, formative assessment results tend to be more detailed than results for other types of assessments, creating the challenge of designing reports or reporting tools that communicate detailed, personalized feedback in ways that are clear and actionable in addition to being automated to support rapid reporting. This may be done by combining digital reporting with assessment designs that provide finer grained results (e.g., Brown et al., 2019) and employing specific strategies such as the use of formative hypotheses to offer suggestions on next steps (Zapata-Rivera et al., 2012) and online tutorials to support comprehension of the reporting tool (Brown et al., 2019). The benefits of feedback to teaching and learning are well documented, and so are the complexities of providing feedback effectively (Hattie & Timperley, 2007; Wisniewski et al., 2020). Technological advances are creating exciting opportunities for advancing the ways we report feedback from formative assessments.

Leveraging Technology for Reporting

Interest in using technology to better communicate assessment results has steadily grown since the start of the millennium. At the time of Goodman and Hambleton's (2004) review of K–12 reporting across 11 U.S. states, only one state posted information related to results reports online. A few years later, this practice had become so common that it prompted the development of guidelines for informational web pages about results reports as well as guidelines for online results-oriented documents and interactive tools (Zenisky & Hambleton, 2012b). Bulut et al. (2020) also proposed recommendations for providing feedback from assessments in an online environment, and Hattie's (2010) commonly cited principles for report development arose from the context of developing web-based, interactive reporting tools.

Many of these recommendations and principles are similar to those for paper-based reports. For instance, Zenisky and Hambleton (2012b) suggested "presenting data in multiple formats when possible" to support users with different information processing needs (p. 181), Hattie (2010) advised that "each report needs to have a major theme" (p. 7), and Bulut et al. (2020) recommended "an aesthetically pleasing design without information overload" (p. 64). Bulut et al.'s (2020) remaining recommendations, presented verbatim, echo several of the guidelines from Figures 13.10 through 13.17.

- The score report should be tailored to meet the needs and characteristics of the target audience, such as students, parents, and teachers (Hambleton & Zenisky, 2013; Zapata-Rivera & Katz, 2014).
- The score report should present the feedback in different forms, including narrative text, tables, and figures.
- The layout of the score report should be simple, with key results highlighted (Goodman & Hambleton, 2004; Slater et al., 2019).

- Feedback presented in the score report should include a set of actions that students can take to improve their future performance (Daniels & Bulut, 2019; Hattie, 2009; Jonsson, 2012).
- If interactive elements (e.g., visuals and tables) are to be used, how students will interact with these elements should be considered in the design process (Bulut et al., 2020; Slater et al., 2019).
- Usability studies with students should be carried out to test whether the content of feedback is easy to follow (Slater et al., 2019; Zenisky & Hambleton, 2012b).

What is different in technology-based reporting is that there is usually no physical constraint on how many elements can be included, and interactivity offers possibilities that were not available before, such as on-demand interpretive text, performance details, and interactive tutorials (e.g., Zapata-Rivera et al., 2016).

In thinking of how to characterize next-generation reports in other innovative ways, it may be that the near future may be app based and use the screen and app-based functions to engage users in ways that current results reports may not (Linares-Vásquez et al., 2017). In the very near future, what users think of as reports may look and feel very different from current "score reports." Other forward-looking research in the areas of big data and visualization may be similarly instructive, in terms of integrating machine learning, developing novel visualizations, conceptualizing different patterns of interaction with data, and alternative user interfaces (Andrienko et al., 2020). This thinking in the area of big data echoes the model-based approach advanced by Zenisky and Hambleton, 2012a), since Andrienko et al. (2020) suggested understanding and "designing the user interactions first" (p. 4), followed by development of systems to support such interactions.

With so many content and design options available, report developers must be very familiar with the needs and preferences of their audience to develop interfaces that will seamlessly guide them—interacting with online resources and tools to explore score data and other results successfully involves greater user engagement than reading a static report. It also may require background knowledge about the assessment itself and the construct measured to formulate specific questions and use the tool or tools as intended. It is also helpful to consider likely misinterpretations and misuses, which become more numerous when interactivity is added, and make adjustments to support the purpose of the assessment. These are critical considerations that developers of online reporting tools must carefully understand and articulate at the outset of that development work, to ensure that these issues are addressed throughout report tool development, including attention paid to providing context and background knowledge for use of any tool. Then, several rounds of gathering user feedback and making refinements are usually needed when developing interactive tools, and relying on a model such as the one proposed by Hambleton and Zenisky (2013) may aid in this process.

Some challenges are unique to online reporting, such as matters of access and ease of navigation. The way users access online reports or reporting tools should be secure

and convenient. Further, the document or tool itself should be easy to navigate, providing users a "guarantee of safe passage" (Hattie, 2010, p. 4), with the most important information made prominent by the design and visual cues, helping users decide where to look next. If the practices of online reporting since the year 2000 are any indication, it is highly likely that online reporting will continue to grow in popularity, bringing about new ways to communicate information from assessments and interesting challenges.

Data Visualization

Approaches to visualizing assessment results have continued to evolve. Classic resources on displaying quantitative data remain relevant (e.g., Kosslyn, 2006; Tufte, 1983, 1990; Wainer, 1997a, 2005) and can now be supplemented by more recent findings from fields such as cognitive science, information visualization, and user experience. Hegarty (2019) provided a thorough discussion of contemporary data visualization best practices and principles that apply specifically to results reporting.

Additionally, the refinement and greater availability of data visualization tools has allowed psychometricians to become more involved in prototyping and producing graphics for reports and reporting tools. Consider the functionality of the following selection of popular open-source data visualization packages for the programming language R (similar tools are also available for Python):

- *ggplot2* (Wickham, 2016): A package for creating static graphics with elegant default features and flexible customization options, based on Wilkinson's (2005) *Grammar of Graphics* structured approach for constructing visualizations (https://ggplot2.tidyverse.org/)
- *plotly* (Sievert, 2020): A package for creating modern, interactive graphics powered by a JavaScript charting library; it may also be used to make ggplot2 graphics interactive using the ggplotly function (https://plotly-r.com/)
- *shiny* (RStudio, 2020): A package and web application framework for creating interactive dashboards with R, often used in conjunction with plotly (https://shiny.rstudio.com/articles/#deployment)

Because results reports are composed of both content and communication, hand in hand with these tools for visualization is an evolution in what reports look like and how the data are accessed by stakeholders. One increasingly common strategy for reporting is the use of digital databases and user-selected reporting queries, and indeed this has shifted the control of the data story from psychometricians and report developers to users (Zenisky & Hambleton, 2012b).

Along the lines of strategies for visualization, a next frontier for reporting, especially with respect to next-generation assessments, is to continue to build on interdisciplinary contributions from cognitive science, marketing and communications, and dissemination practices in other scientific domains to leverage existing and emerging strategies for data understanding and use. In particular, the work of Hegarty (2011,

2019; Padilla et al., 2018), among others (e.g., Hullman et al., 2011; Ratwani et al., 2008), has helped bridge understandings of visual–spatial display cognitions to guide action in a number of areas, such as radiology reports (Alarifi et al., 2021), clinical monitoring (Khasnabish et al., 2020; Reese et al., 2020), meteorology and climate (Argyle et al., 2017; Gerst et al., 2019; Harold et al., 2016), and mapping (Hegarty et al., 2016; Johannsen et al., 2018). There is much to be gained by leveraging modern visualization tools and insights from other fields to improve the visual elements of reports and reporting tools.

Reporting for Next-Generation Assessments

Assessment is changing. From constructs measured to delivery mode, from item formats to sequencing of measurement opportunities in adaptive and gamified assessments, many tests today are substantively different in one or more significant ways than those of just a decade or two ago. As this evolution in tests and measurement continues, it is at least in part driven by users and stakeholders who seek not only different information from assessments but also to use test data more efficiently and effectively (Brown et al., 2023). In this way, reporting—beyond a global score or performance level—is centrally ingrained in the next generation of assessments.

Next-generation assessments are not marked first by technology, but rather a more fundamental shift in paradigm to prioritize process and information, rather than status or outcome. This will necessitate continued work not only on reporting but also next-generation psychometrics, to support the inferences and information that are increasingly sought. The ongoing efforts to report subscores reliably are but a precursor to the challenges that lie ahead in terms of making useful sense of process data (e.g., Bergner & von Davier, 2019; Provasnik, 2021) and, importantly, to devise ways to communicate these data in ways that are supported by evidence. Emerging work in areas such as big data, natural language processing, and digital learning environments will be critical to informing these desired shifts in report contents (Cope & Kalantzis, 2015; Mislevy et al., 2012).

CONCLUSION

> Communicating test score information matters. (Hambleton & Zenisky, 2013, p. 492)

Reporting assessment results is not easy. It is the culmination of a long process of assessment design and development, fundamentally marked by a systematic and serious effort to provide information about individual and/or group performance relative to educational and psychological constructs of interest. As discussed here, reports can be conceptualized through the lens of a data story, where the elements of the report are purposefully chosen, arranged, and field tested with prospective users to communicate specific information about test performance. Much research has focused on reporting in recent years, and the addition of guidance to formalize the process (Hambleton &

Zenisky, 2013) gives structure and purpose to activities that are critical for the development of useful report documents and tools.

In this chapter, the intent has been to reinforce the importance of reporting as part and parcel of the IUA for an assessment. If users are to interpret and use any assessment, they must have access to data in a way that supports those efforts. The work of Kane (2006, 2015) and Hattie (2009, 2010) is instrumental in shaping the perspective on validity and reporting we adopted, and the idea of "interpretability" as suggested by O'Leary et al. (2017) adds an interesting and potentially useful dimension to the conversation around validity and the role of reporting in supporting test interpretation and use.

In some ways, it may almost be easier to focus the conversation about report development on graphics, colors, layout, and design because those are easily manipulated key features of the end product of the report development process. However, as is evident in the earlier conversation around effectiveness, the true test of the success of a report is in its use. The criticality of engaging in reporting needs and wants at the outset of test development cannot be overstated, because the priorities and decisions made early on define the parameters of the data (from a validity perspective) that will be reported later. Interpretations that may be informally desired but are not discussed, defined, and/or prioritized from the beginning are likely to be *unvalidated* and, therefore, relegated to inappropriate uses of results.

If there is a final point to be made, it is the importance of when the conversation about interpretation and use—and reporting—takes place, so that test developers and stakeholder groups alike have a clear understanding of the *what* of reporting, to ensure that the assessment fulfills those known reporting needs (and wants). This takes time, it takes a team, and it takes planning. The evidence for effectiveness (understood as use) comes later, through Phase 4 of the Hambleton and Zenisky model (2012a, 2013) with evaluation and maintenance, but there is no substitute for a strong foundational process for reporting, building on needs assessment, stakeholder input, and consideration of reporting strategies, always, always, in light of test purpose and use.

REFERENCES

Alarifi, M., Patrick, T., Jabour, A., Wu, M., & Luo, J. (2021). Designing a consumer-friendly radiology report using a patient-centered approach. *Journal of Digital Imaging*, *34*(3), 1–12. https://doi.org/10.1007/s10278-021-00448-z

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.

Andrienko, G., Andrienko, N., Drucker, S., Fekete, J. D., Fisher, D., Idreos, S., Kraska, T., Li, G., Ma, K.-L., Mackinlay, J., Oulasvirta, A., Schreck, T., Schumann, H., Stonebreaker, M., Auber, D., Bikakis, N., Chrysanthis, P., Papastefanatos, G., & Sharaf, M. (2020). Big data visualization and analytics: Future research challenges and emerging applications. CEUR Workshop Proceedings, 2578. http://ceur-ws.org/Vol-2578/BigVis1.pdf

Argyle, E. M., Gourley, J. J., Flamig, Z. L., Hansen, T., & Manross, K. (2017). Toward a user-centered design of a weather forecasting decision-support tool. *Bulletin of the American Meteorological Society*, *98*(2), 373–382. https://doi.org/10.1175/BAMS-D-16-0031.1

Bach, B., Stefaner, D., Boy, J., Drucker, S., Bartram, L., Wood, J., Ciuccarelli, P., Engelhardt, Y., Köppen, U., & Tversky, B. (2018). Narrative design patterns for data-driven storytelling. In N. Riche, C. Hurter, N. Diakopoulos, & S. Carpendale (Eds.), *Data-driven storytelling* (pp. 107–133). CRC Press. https://doi.org/10.1201/9781315281575-5

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, *17*(2), 191–204. https://doi.org/10.2307/1165169

Behrens, J. T., DiCerbo, K., Murphy, D., & Robinson, D. (2013, April 28–30). *Conceptual frameworks for reporting results of assessment activities* [Paper presentation]. National Council on Measurement in Education Annual Meeting, San Francisco, CA, United States.

Bennett, E., & Folley, S. (2019). Four design principles for learner dashboards that support student agency and empowerment. *Journal of Applied Research in Higher Education*, *12*(1), 15–26. https://doi.org/10.1108/JARHE-11-2018-0251

Bergner, Y., & von Davier, A. A. (2019). Process data in NAEP: Past, present, and future. *Journal of Educational and Behavioral Statistics*, *44*(6), 706–732. https://doi.org/10.3102/1076998618784700

Bodily, R., Kay, J., Aleven, V., Jivet, I., Davis, D., Xhakaj, F., & Verbert, K. (2018). Open learner models and learning analytics dashboards: A systematic review. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (pp. 41–50). Association for Computing Machinery.

Brennan, R. L. (Ed.). (2006). *Educational measurement* (4th ed.). Praeger.

Brown, G. T. L. (2001). *Reporting assessment information to teachers: Report of Project asTTle outputs design* (asTTle Tech. Rep. #15). University of Auckland. https://www.researchgate.net/publication/237549707_Reporting_Assessment_Information_to_Teachers_Report_of_Project_asTTle_Outputs_Design_Technical_Report_15

Brown, G. T. L., Kannan, P., Sinharay, S., Zapata-Rivera, D., & Zenisky, A. L. (2023). Challenges and opportunities in score reporting: A panel of personal perspectives. *Frontiers in Education*, *8*, 1211580. https://doi.org/10.3389/feduc.2023.121158

Brown, G. T. L., O'Leary, T. M., & Hattie, J. A. C. (2019). Effective reporting for formative assessment: The asTTle case example. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. 107–125). Routledge.

Bulut, O., Cutumisu, M., Singh, D., & Aquilina, A. M. (2020). Guidelines for generating effective feedback from e-assessments. *Hacettepe University Journal of Education*, *35*, 60–72. https://doi:10.16986/HUJE.2020063705

Caldwell, B., Cooper, M., Reid, L. G., & Vanderheiden, G. (2008, December). *Web content accessibility guidelines (WCAG) 2.0*. https://www.w3.org/TR/WCAG20/

Clauser, B. E. (2019, April 4–8). *A history of classical test theory* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Toronto, Canada.

Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring, and reporting. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 355–386). Praeger.

Cope, B., & Kalantzis, M. (2015.) Sources of evidence of learning: Learning and assessment in the era of big data. *Open Review of Educational Research, 2*(1), 194–217. https://doi.org/10.1080/23265507.2015.1074869

Corrin, L., Kennedy, G., French, S., Buckingham Shum S., Kitto, K., Pardo, A., West, D., Mirriahi, N., & Colvin, C. (2019). *The ethics of learning analytics in Australian higher education. A discussion paper*. University of Melbourne CSHE. https://melbourne-cshe.unimelb.edu.au/__data/assets/pdf_file/0004/3035047/LA_Ethics_Discussion_Paper.pdf

Daniels, L. M., & Bulut, O. (2019). Students' perceived usefulness of computerized percentage-only vs. descriptive score reports: Associations with motivation and grades. *Journal of Computer Assisted Learning, 36*(2), 199–208. https://doi:10.1111/jcal.12398

De Boeck, P., & Scalise, K. (2019). Collaborative problem solving: Processing actions, time, and performance. *Frontiers in Psychology, 10*, Article 1280. https://doi.org/10.3389/fpsyg.2019.01280

Du, X., Yang, J., Shelton, B. E., Hung, J.-L., & Zhang, M. (2021) A systematic meta-review and analysis of learning analytics research. *Behaviour & Information Technology*, *40*(1), 49–62. https://doi.org/10.1080/0144929X.2019.1669712

Ercikan, K., & Pellegrino, J. W. (Eds.). (2017). *Validation of score meaning for the next generation of assessments: The use of response processes*. Routledge.

Ferrara, S., & Lai, E. (2016). Documentation to support test score interpretation and use. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 603–623). Routledge.

Few, S. (2006). *Information dashboard design: The effective visual communication of data*. O'Reilly Media.

Fischman, G. E., Topper, A. M., Silova, I., Goebel, J., & Holloway, J. L. (2019.) Examining the influence of international large-scale assessments on national education policies. *Journal of Education Policy*, *34*(4), 470–499. https://doi.org/10.1080/02680939.2018.1460493

Forte Fast, E., & Tucker, C. (2001, April 10–14). *Redesign of the student assessment reporting system in Connecticut* [Paper presentation]. American Educational Research Association Annual Meeting, Seattle, WA, United States.

Gerst, M. D., Kenney, M. A., Baer, A. E., Speciale, A., Wolfinger, J. F., Gottschalck, J., Handel, S., Rosencrans, M., & Dewitt, D. (2019). Using visualization science to improve expert and public understanding of probabilistic temperature and precipitation outlooks. *Weather, Climate, and Society*, *12*(1), 117–133. https://doi.org/10.1175/WCAS-D-18-0094.1

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145–220. https://doi.org/10.1207/s15324818ame1702_3

Gotch, C. M., & Roduta Roberts, M. (2018). A review of recent research on individual-level score reports. *Educational Measurement: Issues and Practice, 37*, 46–54. https://doi.org/10.1111/emip.12198

Hambleton, R. K. (1998). Enhancing the validity of NAEP achievement level score reporting. In M. L. Bourque (Ed.), *Proceedings of the Achievement Levels Workshop* (pp. 77–98). National Assessment Governing Board.

Hambleton, R. K. (2002). How can we make NAEP and state test score reporting scale and reports more understandable? In R. W. Lissitz & W. D. Schafer (Eds.), *Assessment in educational reform* (pp. 192–205). Allyn & Bacon.

Hambleton, R. K., & Meara, K. (2000). Newspaper coverage of NAEP results, 1990–1998. In M. L. Bourque & S. Byrd (Eds.), *Student performance standards on the National Assessment of Educational Progress: Affirmation and improvements. A study initiated to examine a decade of achievement level setting on NAEP* (pp. 133–155). National Assessment Governing Board.

Hambleton, R. K., & Slater, S. C. (1997). *Are NAEP executive summary reports understandable to policy makers and educators?* (CSE Technical Report 430). National Center for Research on Evaluation, Standards, and Student Teaching.

Hambleton, R. K., & Zenisky. A. L. (2003). Advances in criterion-referenced testing methods and practices. In C. R. Reynolds & R. W. Kamphaus (Eds.), *The handbook of psychological and educational assessment* (2nd ed., pp. 377–404). The Guilford Press.

Hambleton, R. K., & Zenisky, A. L. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology. Vol. 3: Testing and assessment in school psychology and education* (pp. 479–494). American Psychological Association. https://doi.org/10.1037/14049-023

Hambleton, R. K., & Zenisky, A. L. (2017). Score reporting and interpretation. In W. J. van der Linden (Ed.), *Handbook of modern item response theory* (2nd ed., pp. 127–142). Chapman and Hall/CRC.

Hamilton, L. S., Stecher, B. M., & Yuan, K. (2009.) *Standards-based reform in the United States: History, research, and future directions*. Center on Education Policy.

Harold, J., Lorenzoni, I., Shipley, T., & Coventry, K. (2016). Cognitive and psychological science insights to improve climate change data visualization. *Nature Climate Change, 6*, 1080–1089. https://doi.org/10.1038/nclimate3162

Hattie, J. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hattie, J. (2010). Visibly learning from reports: The validity of score reports. *Online Educational Research Journal, 1*(5), 1–15. https://oerj.webspace.durham.ac.uk/wp-content/uploads/sites/249/2021/07/4.pdf

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hegarty, M. (2011). The cognitive science of visual–spatial displays: Implications for design. *Topics in Cognitive Science, 3*, 446–474. https://doi.org/10.1111/j.1756-8765.2011.01150.x

Hegarty, M. (2019). Advances in cognitive science and information visualization. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. 19–34). Routledge.

Hegarty, M., Friedman, A., Boone, A., & Barrett, T. J. (2016). Where are you? The effect of uncertainty and its visual representation on location judgments in GPS-like displays. *Journal of Experimental Psychology: Applied, 22*(4), 381–392. https://doi.org/10.1037/xap0000103

Hooper, L. (2021). *How to tell a story with data: A guide for beginners*. Venngage. https://venngage.com/blog/data-storytelling/

Hullman, J., Rhodes, R., Rodriguez, F., & Shah, P. (2011, November). *Research on graph comprehension and data interpretation: Implications for score reporting* [Paper presentation]. ETS Score Reporting Conference, Princeton, NJ, United States.

International Test Commission. (2013). *International guidelines for test use* (version 1.2). Retrieved March 2, 2024, from https://www.intestcom.org/files/guideline_test_use.pdf

Jaeger, R. M. (2003). *NAEP validity studies: Reporting the results of the National Assessment of Educational Progress* (Working Paper 2003-11). U.S. Department of Education, Institute of Education Sciences.

Jaeger, R.M., Gorney, B., Johnson, R. L., Putnam, S. E., & Williamson, G. (1993). *Designing and developing effective school report cards: A research synthesis*. Center for Research on Education Accountability and Teacher Evaluation, Western Michigan University.

Johannsen, I. M., Lassonde, K. A., Wilkerson, F., & Schaab, G. (2018). Communicating climate change: Reinforcing comprehension and personal ties to climate change through maps. *The Cartographic Journal, 55*(1), 85–100. https://doi.org/10.1080/00087041.2017.1386834

Jonsson, A. (2012). Facilitating productive use of feedback in higher education. *Active Learning in Higher Education, 14*(1), 63–76. https://doi.org/10.1177/1469787412467125

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.

Kane, M. T. (2015). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 64–80). Routledge.

Kannan, P., & 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. *Frontiers in Education, 7*, 913594. https://doi.org/10.3389/feduc.2022.913594

Kannan, P., Zapata-Rivera, D., & Leibowitz, E. A. (2018). Interpretation of score reports by diverse subgroups of parents. *Educational Assessment, 23*(3), 173–194. https://doi.org/10.1080/10627197.2018.1477584

Katz, I. R. (2019). Foreword. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. xiii–xiv). Routledge.

Kay, J. (1997). Learner know thyself: Student models to give learner control and responsibility. In Z. Halim, T. Ottomann, & Z. Razak (Eds.), *Proceedings of*

International Conference on Computers in Education (pp. 17–24). AACE. https://www1.inhatc.ac.kr/seihoon/papers/learning/Kay97.pdf

Khasnabish, S., Burns, Z., Couch, M., Mullin, M., Newmark, R., & Dykes, P. C. (2020). Best practices for data visualization: Creating and evaluating a report for an evidence-based fall prevention program. *Journal of the American Medical Informatics Association*, *27*(2), 308–314. https://doi.org/10.1093/jamia/ocz190

Kirylo, J. D. (2018). The opt-out movement and the power of parents. *Phi Delta Kappan*, *99*(8), 36–40. https://doi.org/10.1177/0031721718775676

Knupp, T., & Ansley, T. (2008, March 25–27). *Online, state-specific assessment score reports and interpretive guides* [Paper presentation]. National Council on Measurement in Education Annual Meeting, New York, NY, United States.

Kosslyn, S. M. (2006). *Graph design for the eye and mind*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195311846.001.0001

Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. Farrar, Straus and Giroux.

Levine, R., Rathbun, A., Selden, R., & Davis, A. (1998). *NAEP's constituents: What do they want? Report of the National Assessment of Educational Progress constituents survey and focus groups* (NCES 98–521). U.S. Department of Education, Office of Educational Research and Improvement.

Linares-Vásquez, M., Moran, K., & Poshyvanyk, D. (2017, September). Continuous, evolutionary and large-scale: A new perspective for automated mobile app testing. In *2017 IEEE International Conference on Software Maintenance and Evolution* (pp. 399–410). IEEE.

Lindquist, E. F. (Ed.). (1951). *Educational measurement*. American Council on Education.

Linn, R. L. (Ed.). (1989). *The American Council on Education/Macmillan series on higher education. Educational measurement* (3rd ed.). American Council on Education and Macmillan.

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher*, *31*(6), 3–16. https://doi.org/10.3102/0013189X031006003

Luecht, R. (2013). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. *Journal of Applied Testing Technology*, *14*, 1–38. https://www.testpublishers.org/jatt-volume-14

MacIver, R., Anderson, N., Costa, A. C., & Evers, A. (2014). Validity of interpretation: A user validity perspective beyond the test score. *International Journal of Selection and Assessment*, *22*(2), 149–164. https://doi.org/10.1111/ijsa.12065

Measurement Research Associates, Inc. (2010). *Using the very useful Wright map*. Retrieved August 26, 2015, from http://www.rasch.org/mra/mra-01-10.htm

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education.

Milligan, S. K., & Griffin, P. (2016). Understanding learning and learning design in MOOCs: A measurement-based interpretation. *Journal of Learning Analytics*, *3*(2), 88–115.

Mislevy, R. J. (1998). Implications of market-basket-reporting for achievement level setting. *Applied Measurement in Education, 11*(1), 49–63. https://doi.org/10.1207/s15324818ame1101_3

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining, 4*(1), 11–48. https://doi.org/10.5281/zenodo.3554641

Mislevy, R., & Haertel, G. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*, 6–20. https://doi.org/10.1111/j.1745-3992.2006.00075.x

Mosier, C. I. (1951). Batteries and profiles. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 764–808). American Council on Education.

National Center for Education Statistics. (2013). *The Nation's Report Card: Trends in academic progress 2012* (NCES 2013–456). Institute of Education Sciences, U.S. Department of Education.

National Commission for Certifying Agencies. (2014). *Standards for the accreditation of certification programs*. Institute for Credentialing Excellence.

National Commission on Excellence in Education. (1983). *A Nation at Risk: The imperative for educational reform*. U.S. Government Printing Office.

National Research Council. (2001). *NAEP reporting practices: Investigating district-level and market-basket reporting*. National Academies Press.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 101, Stat. 1425 (2002).

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw–Hill.

O'Donnell, F. (2020). *What's in a label? Unpacking the meaning of achievement labels from tests* [Doctoral dissertation, University of Massachusetts Amherst]. 1856. https://doi.org/10.7275/15558737

O'Donnell, F., Ong, T. Q., & Feinberg, R. (2021, May 18–June 11). Advances in reporting results on medical education assessments. In C. Runyon & S. Somay (Chairs), *Advancing assessment in medical education* [Coordinated session]. National Council on Measurement in Education Annual Meeting.

O'Donnell, F., & Sireci, S. G. (2019). Score reporting issues for licensure, certification, and admissions programs. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. 77–90). Routledge.

O'Donnell, F., & Sireci, S. G. (2021). Language matters: Teacher and parent perceptions of achievement labels from educational tests. *Educational Assessment, 27* (1), 1–26. https://doi.org/10.1080/10627197.2021.2016388

O'Donnell, F., & Zenisky, A. L. (2020), Digital module 21: Results reporting for large-scale assessments. *Educational Measurement: Issues and Practice, 39*, 137–138. https://doi.org/10.1111/emip.12408

O'Leary, T. M., Hattie, J. C., & Griffin, P. (2017). Actual interpretations and use of scores as aspects of validity. *Educational Measurement: Issues & Practice, 36*(2), 16–23. https://doi:10.1111/emip.12141

Oliva-Córdova, L. M., Garcia-Cabot, A., & Amado-Salvatierra, H. R. (2021). Learning analytics to support teaching skills: A systematic literature review. *IEEE Access, 9*, 58351–58363. https://doi.org/10.1109/ACCESS.2021.3070294

Padilla, L. M., Creem-Regehr, S. H., Hegarty, M., & Stefanucci, J. K. (2018). Decision making with visualizations: A cognitive framework across disciplines. *Cognitive Research: Principles and Implications, 3*(1), 29–54. https://doi.org/10.1186/s41235-018-0120-9

Park, Y., & Jo, I. (2019). Factors that affect the success of learning analytics dashboards. *Educational Technology Research and Development, 67*(6), 1547–1571. https://doi.org/10.1007/s11423-019-09693-0

Provasnik, S. (2021). Process data, the new frontier for assessment development: Rich new soil or a quixotic quest? *Large-Scale Assessments in Education, 9*(1), 17–17. https://doi.org/10.1186/s40536-020-00092-z

Ratwani, R. M., Trafton, J. G., & Boehm-Davis, D. A. (2008). Thinking graphically: Connecting vision and cognition during graph comprehension. *Journal of Experimental Psychology: Applied, 14*(1), 36–49. https://doi.org/10.1037/1076-898X.14.1.36

Reese, T. J., Segall, N., Del Fiol, G., Tonna, J. E., Kawamoto, K., Weir, C., & Wright, M. C. (2020). Iterative heuristic design of temporal graphic displays with clinical domain experts. *Journal of Clinical Monitoring and Computing, 35*(5), 1119–1131. https://doi.org/10.1007/s10877-020-00571-2

Roberts, M., & Gierl, M. (2010). Developing score reports for cognitive diagnostic assessments. *Educational Measurement: Issues and Practice, 29*(3), 25–38. https://doi.org/10.1111/j.1745-3992.2010.00181.x

Roduta Roberts, M., Gotch C. M., & Lester, J. N. (2018). Examining score report language in accountability testing. *Frontiers in Education, 3*(42), 1–17. https://doi.org/10.3389/feduc.2018.00042

RStudio. (2020). *Shiny: Easy web applications in R*. https://shiny.rstudio.com/

Ryan, J. M. (2006). Practices, issues, and trends in student test score reporting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 677–710). Lawrence Erlbaum Associates.

Sarikaya, A., Correll, M., Bartram, L., Tory, M., & Fisher, D. (2019). What do we talk about when we talk about dashboards? *IEEE Transactions on Visualization and Computer Graphics, 29*(1), 682–692.

Schwendimann, B., Rodriguez-Triana, M., Vozniuk, A., Prieto, L., Boroujeni, M., Holzer, A., Gillet, D., & Dillenbourg, P. (2017). Perceiving learning at a glance: A systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies, 10*(1), 30–41. https://doi.org/10.1109/tlt.2016.2599522

Siemens, G., & Baker, R. S. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 252–254). ACM.

Sievert, C. (2020). *Interactive web-based data visualization with R, plotly, and shiny*. Chapman & Hall/CRC. https://plotly-r.com

Simmons, C., & Mwalimu, M. (2000). What NAEP's publics have to say. In M. L. Bourque & S. Byrd (Eds.), *Student performance standards on the National Assessment of Educational Progress: Affirmation and improvements. A study initiated to examine a decade of achievement level setting on NAEP* (pp. 184–219). National Assessment Governing Board.

Sinharay, S., Haberman, S. A., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice, 26*(4), 21–28.

Sinharay, S., Puhan, G., Haberman, S. J., & Hambleton, R. K. (2019). Subscores: When to communicate them, what are their alternatives, and some recommendations. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. 35–49). Routledge.

Slater, S., Livingston, S. A., & Silver, M. (2019). Score reports for large-scale testing programs: Managing the design process. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. 91–106). Routledge.

Tannenbaum, R. J. (2019). Validity aspects of score reporting. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. 9–18). Routledge.

Thorndike, R. (Ed.). (1971). *Educational measurement* (2nd ed.). American Council on Education.

Thudt, A., Gschwandtner, T., Walny, J., Dykes, J. & Stasko, J. (2018). Exploration and explanation in data-driven storytelling. In N. Riche, C. Hurter, N. Diakopoulos, & S. Carpendale (Eds.), *Data-driven storytelling* (pp. 59–83). CRC Press. https://doi.org/10.1201/9781315281575-5

Tufte, E. R. (1983). *The visual display of quantitative information*. Graphics Press.

Tufte, E. R. (1990). *Envisioning information*. Graphics Press.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. (2015). *NAEP science—map of selected item descriptions on the NAEP Science Scale—Grade 8*. Retrieved August 25, 2015, from http://nces.ed.gov/nationsreportcard/itemmaps/index.asp

Van der Kleij, F. M., Eggen, T. J. H. M., & Engelen, R. J. H. (2014). Towards valid score reports in the computer program LOVS: A redesign study. *Studies in Educational Evaluation, 43*, 24–39. https://doi.org/10.1016/j.stueduc.2014.04.004

Verbert, K., Govaerts, S., Duval, E., Santos, J., Van Assche, F., Parra, G., & Klerkx, J. (2013). Learning dashboards: An overview and future research opportunities. *Personal and Ubiquitous Computing, 18*(6), 1499–1514. https://doi.org/10.1007/s00779-013-0751-2

Wainer, H. (1992). Understanding graphs and tables. *Educational Researcher, 21*(1), 14–23. https://doi.org/10.3102/0013189X021001014

Wainer, H. (1997a). *Visual revelations: Graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot*. Copernicus Books.

Wainer, H. (1997b). Improving tabular displays: With NAEP tables as examples and inspirations. *Journal of Educational and Behavioral Statistics, 22*(1), 1–30. https://doi.org/10.3102/10769986022001001

Wainer, H. (2005). *Graphic discovery: A trout in the milk and other visual adventures*. Princeton University Press.

Wainer, H. (2009). *Picturing the uncertain world: How to understand, communicate, and control uncertainty through graphical display*. Princeton University Press.

Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement, 36*(4), 301–335. https://doi.org/10.1111/j.1745-3984.1999.tb00559.x

Waters, B. K. (1997). Army Alpha to CAT-ASVAB: Fourscore years of military personnel selection and classification testing. In R. F. Dillon (Ed.), *Handbook on testing* (pp. 187–203). Greenwood Press.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

Wilkinson, L. (2005). *The grammar of graphics* (2nd ed.). *Statistics and computing*. Springer.

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*, 1–15. https://doi.org/10.3389/fpsyg.2019.03087

Xiao, Y., Veldkamp, B., & Liu, H. (2022), Combining process information and item response modeling to estimate problem-solving ability. *Educational Measurement: Issues and Practice, 41*, 36–54. https://doi.org/10.1111/emip.12474

Yen, W. M. (1987). *A Bayesian/IRT index of objective performance* [Paper presentation]. Meeting of the Psychometric Society, Montreal, Canada.

Zapata-Rivera, D. (2011). Designing and evaluating score reports for particular audiences. In D. Zapata-Rivera & R. Zwick (Eds.), *Test score reporting: Perspectives from the ETS Score Reporting Conference* (ETS Research Report No. RR-11-45). ETS. https://www.ets.org/Media/Research/pdf/RR-11-45.pdf

Zapata-Rivera, D. (2021). Open student modeling research and its connections to educational assessment. *International Journal of Artificial Intelligence in Education, 31*, 380–396. https://doi.org/10.1007/s40593-020-00206-2

Zapata-Rivera, D., Andrews-Todd, J., & Oliveri, M. E. (2021). Communicating assessment information in the context of a workplace formative task. *The Journal of Writing Analytics, 5*, 324–341. https://doi.org/10.37514/JWA-J.2021.5.1.10

Zapata-Rivera, D., Kannan, P., & Zwick, R. (2019). Communicating measurement error information to teachers and parents. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. 63–73). Routledge.

Zapata-Rivera, D., & Katz, R. I. (2014). Keeping your audience in mind: Applying audience analysis to the design of interactive score reports. *Assessment in Education: Principles, Policy & Practice, 21*(3), 442–463.

Zapata-Rivera, D., VanWinkle, W., & Zwick, R. (2012). *Applying score design principles in the design of score reports for CBAL™ teachers* (ETS Research Memorandum RM-12-20). ETS.

Zapata-Rivera, D., Zwick, R., & Vezzu, M. (2016). Exploring the effectiveness of a measurement error tutorial in helping teachers understand score report results.

Educational Assessment, 21(3), 215–229. https://doi.org/10.1080/10627197.2016.1202110

Zenisky, A. L. (2015). Visual displays for reporting test data: Making sense of test performance. In M. McCrudden, G. Schraw, & C. Buckendahl (Eds.), *Use of visual displays in research and testing: Coding, interpreting, and reporting data* (pp. 299–332). Information Age Publishing.

Zenisky, A. L., Delton, J., & Hambleton, R. K. (2006). *State reading content specialists and NAEP data displays* (Report No. 598). University of Massachusetts, Center for Educational Assessment.

Zenisky, A. L., & Hambleton, R. K. (2007). *Navigating 'The Nation's Report Card" on the World Wide Web: Site user behavior and impressions* [Technical report]. Comprehensive Evaluation of NAEP. [Also Center for Educational Assessment Report No. 625]

Zenisky, A. L., & Hambleton, R. K. (2012a). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice, 31*(2), 21–26. https://doi.org/10.1111/j.1745-3992.2012.00231.x

Zenisky, A. L., & Hambleton, R. K. (2012b). From "here's the story" to "you're in charge": Developing and maintaining large-scale online test and score reporting resources. In M. Simon, M. Rousseau, & K. Ercikan (Eds.), *Improving large-scale assessment in education* (pp. 175–185). Routledge.

Zenisky, A. L., & Hambleton, R. K. (2015). Test score reporting: Best practices and issues. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed, pp. 585–602). Routledge.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2009). Getting the message out: An evaluation of NAEP score reporting practices with implications for disseminating test results. *Applied Measurement in Education, 22*(4), 359–375. https://doi.org/10.1080/08957340903221667

Zenisky, A. L., Hambleton, R. K., & Smith, Z. R. (2006). *Do math educators understand NAEP score reports? Evaluating the utility of selected NAEP data displays* (Report No. 587). University of Massachusetts, Center for Educational Assessment.

Zenisky, A. L., Keller, L. A., & Park, Y. (2019). Reporting student growth: Challenges and opportunities. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. 50–62). Routledge.

Zenisky, A. L., Mazzeo, J., & Pitoniak, M. J. (2016). Towards improving the reporting of test score results: Lessons learned from the evolution of NAEP reports. In. C. Wells & M. Faulkner-Bond (Eds.), *Educational measurement: From foundations to future* (pp. 335–355). Guilford.

Zenisky, A. L., Sireci, S. G., Lewis, J., Lim, H., O'Donnell, F., Wells, C. S., Padellaro, F., Jung, H. J., Banda, E., Pham, D., Hong, S. E., Park, Y., Botha, S., Lee, M., & Garcia, A. (2018). *Massachusetts adult proficiency tests for college and career readiness: Technical manual* (Center for Educational Assessment Research Report No. 974). Center for Educational Assessment.

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa*, *20*, 79–87. https://doi.org/10.1016/j.pse.2014.11.003

Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error in test score reports. *Educational Assessment*, *19*(2), 116–138. https://doi.org/10.1080/10627197.2014.903653.

NOTE

1. Throughout this chapter, we honor the legacy of our colleague and this chapter's third author, Ronald K. Hambleton, by incorporating specific quotations drawn from his extensive body of work on this topic. Professor Hambleton's countless contributions have significantly shaped our field and continue to frame our discourse. His passing on April 28, 2022, was a profound loss, but his voice and influence endure through his writings, which we are privileged to share here.