9

Technology-Based Assessment

Validity, Modeling, and Analysis Issues

Randy E. Bennett
ETS

Michelle LaMar
ETS (Retired)

John Mazzeo
ETS (Retired)

Bennett, R. E., LaMar, M., & Mazzeo, J. (2025). Technology-based assessment: Validity, modeling, and analysis issues. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 581–654). Oxford University Press.
DOI: 10.1093/oso/9780197654965.003.0009

In this chapter, we focus on validity, modeling, and analysis issues in technology-based assessment. For present purposes, we define *technology-based assessment* (TBA) as a measurement used for decision-making primarily in education, but also in the workplace, that employs digital computing in most, if not all, aspects of its creation, delivery, presentation, scoring, or reporting.

As of this writing, TBAs were being used for consequential decision-making purposes widely in the United States, as well as in some other countries and in some international assessments. Among the more prominent of such programs are the National Assessment of Educational Progress (NAEP), the Programme for International Student Assessment (PISA), the Australian National Assessment Program, the GRE, the TOEFL, the Graduate Management Admission Test, the United States Medical Licensing Examination (USMLE), the Law School Admission Test, and the California Assessment of Student Performance and Progress (CAASPP). CAASPP is particularly notable because of its test-taker volume, which may be the largest of any such measure. On a single day, June 7, 2019, that volume exceeded 670,000 students, with well over 3 million individuals taking the examinations in the spring 2019 cycle (Johnson, 2019). Although many testing programs, especially those outside the United States, continue to test on paper, the number of major testing programs that have become digital, coupled with the large size of CAASPP, demonstrates that technology delivery is feasible at scale. The transition has proven to be, as one of the authors of this chapter had earlier suggested, "inevitable and inexorable" (Bennett, 2002).

What is motivating this transition? Briefly stated, the most salient reasons revolve around three areas. First is the need to align the medium of testing with that of learning and of the information economy's workplace. Absent that alignment, assessment runs the risk of appearing, and becoming, irrelevant to its constituents. The second reason is the belief that assessment processes can be conducted more efficiently, with savings in time and cost. Scores, for example, can in some instances be generated and reported immediately. Third is the measurement of constructs that are impossible to evaluate in traditional testing modes. Examples include using technology tools for such activities as reading in hyperlinked environments, information search and synthesis, writing, modeling, and collaborative problem-solving (Institute for Education Sciences, n.d., 2018; Mullis & Prendergast, 2017).

The transition that has occurred since the 1990s in the United States and elsewhere can be described at a high level in terms of three stages or generations (Bennett, 1998, 2010a). The first generation is essentially an infrastructure-building effort. The cost and time required to put that infrastructure into place are typically substantial, involving obtaining hardware and software, hiring personnel, providing training, and creating myriad new processes and procedures. To control cost and complication, tests in this generation often look little different from paper assessments in their realized design and question format. This generation's tests primarily serve institutional purposes like school accountability, leverage technology minimally (e.g., through incremental advances like adaptive testing), and are generally organized as singular events (e.g., for state accountability purposes, as annual, end-of-year administrations).

In the second generation, goals begin to shift toward achieving some degree of qualitative change and to leveraging the technology more effectively for efficiency improvement (Bennett, 1998, 2010a). With respect to the former goal, a diversity of less traditional item formats may be employed, including those sometimes called "technology-enhanced" (e.g., involving multimedia stimuli, dragging and dropping onscreen text or objects, highlighting segments of text). Additionally, new constructs may be introduced (e.g., writing on computer), so that what is assessed begins to change fundamentally. Efficiency improvements extend beyond delivery via automated approaches to item generation (Gierl & Haladyna, 2013; Irvine & Kyllonen, 2010), test assembly (Veldkamp & Paap, 2017), and scoring (Shermis & Burstein, 2013; see also Shermis et al., this volume), as well as to using the Internet for such processes as item review, standard setting, human online scoring, reporting, and other communications with test users.

Reinvention characterizes the third generation of assessments (Bennett, 1998, 2010a). What in the two earlier generations was an evolution dictated primarily by technology now shifts to one driven by substance. In this generation, theory-based models and cognitive principles combine with more traditional content considerations to provide the substantive basis for assessment design (see Huff et al., this volume). Second, these assessments integrate the needs of individuals more fully with those of institutions. A special case of this development is greater integration with instruction, including the repeated sampling of performance over time. Third and finally, the use of complex simulations and other interactive performance tasks allows new skills to be measured and traditional ones to be evaluated in more meaningful ways.

Of note with respect to the focus of this chapter is that the *Standards for Educational and Psychological Testing* (*Standards*; American Educational Research Association [AERA] et al., 2014) contains no section or set of standards specifically devoted to technology-based assessment. The document does, however, make clear that "computer . . . tests need to be held to the same requirements of technical quality as other tests" (p. 197) and that the "interpretation of scores on technology-based tests are evaluated by the same standards for validity, reliability/precision, and fairness as tests administered through more traditional means" (p. 188). In this sense, an assessment's purpose, with its attendant claims for score interpretation and use, remains the central consideration in determining the appropriate level of technical quality.

With the above as an introduction, the goals of this chapter are to present (a) major issues in validity, modeling, and analysis for TBAs and (b) potential approaches to addressing those issues. The chapter is organized as follows. The first section centers on assessments used to support consequential purposes. This class encompasses decisions that, as singular instances, may have highly significant impact on individuals, groups, or institutions and that are often difficult to reverse. Examples include school admissions, promotion and graduation, educator evaluation, school accountability, intranational and international comparisons, and job licensure and certification. In this section, the history and current landscape are described for technology-based tests of this type.

Discussed are innovative tasks and item types, comparability, the analysis of response processes, and the automated scoring of complex constructed-response tasks.

The second major section covers assessments employed for in-the-moment instructional decisions or for describing what a student knows and can do so that near-term (but not necessarily real-time) instructional next steps can be taken. As singular instances, these decisions usually have less dramatic impact on individuals and are more easily reversed than decisions we have termed *consequential*. This section also begins with a review of the history and current landscape that includes brief consideration of common technology-based assessment designs and task types, response processes, and scoring methods. This section closes with a discussion of psychometric models for assessment embedded in instructional and learning systems.

The chapter's last major section explores the third-generation idea of combining both assessment purposes—that is, consequential decision-making and instructional support—in the same assessment. We conclude the chapter by summarizing key points, giving recommendations for research, and speculating on future directions.

ASSESSMENTS USED TO SUPPORT CONSEQUENTIAL PURPOSES

In this section, we outline the landscape regarding TBAs used for consequential purposes in education, as well as in the professions and occupations. We include tests used to make consequential decisions about individual test takers, as well as state, federal, and international measures intended to provide results for various monitoring and accountability purposes. In doing so, we trace major milestones in the development and operational deployment of TBAs.

Even prior to the widespread availability of desktop computers, the potential benefits of testing by computer were relatively obvious. Such delivery offered the possibility of immediate scoring and reductions in cost associated with the elimination of printing and shipping of test materials. Throughout the 1970s and into the 1980s, advances in psychometric research and theory by Lord (1970) and Weiss (1976) revealed how more powerful psychometric models, coupled with alternative test delivery schemes enabled by computer, could also improve quality. Computer delivery that moved beyond simple, linear, preassembled test forms to the use of sequential and computer-adaptive tests (CAT) offered the potential of increased measurement precision throughout the proficiency range, as well as increased efficiency (i.e., better precision for a fixed amount of testing time; Lord, 1980; Weiss, 1982).

The operationalization of TBA required an infrastructure to deliver tests securely on a large scale. As described by W. D. Way and Robin (2016), this infrastructure first emerged, surprisingly, from the University of Illinois's efforts to expand computer-assisted instruction (CAI) as implemented in its Programmed Logic for Automatic Teacher Operations system. That system was further developed and commercialized by Control Data Corporation. In partnership with the National

Association of Securities Dealers, the Control Data Corporation used this system, circa 1978, to offer what was most likely the first proctored TBA for consequential decisions, in this instance for licensure and certification.

One of the earliest entities interested in the potential convenience and efficiency of TBA—in particular, through adaptive testing—was the U.S. military (Sands et al., 1997). Military institutions sponsored considerable research and influential conferences on this topic that advanced the measurement field substantially (Weiss, 1978, 1980, 1985). In 1979, a feasibility study was launched to develop an adaptive version of the military's selection and placement test battery, the Armed Services Vocational Aptitude Battery. This test was developed, evaluated, launched for operational testing in 1992, and adopted fully in 1996.

One of the earliest efforts to deploy TBAs for consequential decisions in education—in this case, college placement decisions—began in the mid-1980s, when the College Board introduced its computerized-adaptive College Placement Tests (CPTs). The CPTs, delivered on microcomputers, were intended for use at 2- and 4-year institutions to help identify incoming students requiring remedial education in English, reading, and mathematics. The development and validation of these tests is described by Ward (1988) and Ward et al. (1986). The College Board's current college placement test—Accuplacer—is a direct descendant of the CPTs.

A subsequent major milestone was the development and introduction by ETS (Educational Testing Service) in 1993 of a CAT version of the GRE General Test, for use in graduate admissions decisions (Mills, 1999). Designed to award scores that were comparable to its paper-and-pencil predecessor, this work produced a key innovation: an item-selection algorithm capable of balancing psychometric considerations with constraints related to item content, format, and exposure, including dealing with passage-based items (Stocking & Swanson, 1993). This innovation helped to ensure that GRE CAT scores not only were psychometrically efficient but also met the content validity and fairness criteria important to consequential use. In this same decade, ETS followed with computer versions for other postsecondary admissions programs, including the adaptive Graduate Management Admission Test and the TOEFL CBT, with both adaptive and linear sections.

Whereas the introduction of these TBAs was a notable advance, it also surfaced some unanticipated challenges (Wainer & Eignor, 2000). Paper-and-pencil tests could be administered to large numbers of students on a few days per year, requiring only a small number of different forms. In contrast, because of infrastructure limits, TBAs had to be delivered in small centers to fewer students on a more continuous basis. Since demand tended to peak at particular times, problems quickly arose with test-taker access to centers. As important, it soon became apparent that initial plans for item pool sizes and item exposure controls were inadequate to maintain security. Dealing effectively with these problems substantially increased the costs of CAT TBA relative to paper-and-pencil administration. Additional problems emerged with respect to preventing students from learning how to "game" the CAT (W. D. Way & Robin, 2016). The vast

majority of these problems have been ameliorated over time, with increased access, use of alternative models such as multistage adaptive testing (MST), and enhanced schemes for exposure control. As of this writing, the two major undergraduate college admissions programs (ACT, formerly American College Testing; and SAT, formerly Scholastic Aptitude Test) have begun to offer TBA versions, with the SAT having become all digital as of 2024 (College Board, 2024).

The first forays into TBA consisted largely of adaptively delivering multiple-choice items, the staple of paper-and-pencil testing programs. Relatively little was done to leverage the affordances of TBAs for presenting more complex tasks focused on measuring a broader set of competencies. While this eventuality may have been partly due to the limitations of the available computer technology, other contributing factors included cost, limits on available testing time, score reliability, comparability with paper versions, and fairness concerning the use of performance testing generally.

It is perhaps not surprising, then, that early attempts to introduce performance tasks occurred with licensure and certification testing, where longer testing times, costs that could be passed on to the test taker, and reduced need to maintain comparability with prior versions made such innovations more tractable. For example, in the 1990s the National Council of Architectural Registration Boards introduced an automatically scored, performance-based design section into its licensure examination (Bejar & Braun, 1999). Similarly, the National Board of Medical Examiners incorporated computer-based patient management cases into the USMLE (Margolis & Clauser, 2006). Finally, the American Institutes of Certified Public Accountants began to administer TBAs containing performance tasks (Breithaupt et al., 2006). It is noteworthy that all these examinations retained very substantial, complementary multiple-choice sections to achieve the reliability and generalizability levels needed for licensure decisions. (See Margolis et al., this volume, for a comprehensive review of testing in licensure and certification.)

The first waves of consequential TBA testing—in the military, in university placement, in graduate admissions, and in licensure and certification—occurred with TBAs delivered on mainframes, then on stand-alone desktop computers in college placement offices, and next via networked machines at vendor testing centers (e.g., Prometric and Pearson VUE) and in university labs. Hardware and software could be reasonably standardized, and the use of data networks to transmit test content and response data allowed security to be maintained. For K–12 education, in contrast, the use of TBA for state-mandated accountability tests, high school end-of-course measures, and graduation examinations awaited development of a more pervasive infrastructure that could accommodate machines resident in the schools. (See Ho & Polikoff, this volume, for a comprehensive discussion of assessment for accountability in K–12.)

With the evolution of the Internet and the proliferation of laptops and tablets, TBA became increasingly practical for state-mandated accountability tests. This history dates to about 2000, when Oregon, Virginia, and a few other states each began pilot TBA programs (Bennett, 2002). Those pilot efforts gradually expanded so that by

2007, Oregon, for example, had received approval for a computer-adaptive version of its federally mandated K–8 summative assessment (W. D. Way & Robin, 2016). The passage of the American Recovery and Reinvestment Act of 2009, and in particular the Race to the Top Assessment Program (U.S. Department of Education, n.d.), gave a dramatic boost to development by funding several testing consortia, including the Smarter Balanced Assessment Consortium and the Partnership for the Assessment of Readiness for College and Careers (PARCC). Smarter Balanced and PARCC developed TBAs aligned to the Common Core State Standards in English language arts and mathematics. The tests made use of adaptivity (Smarter Balanced), items that employed technology affordances (technology-enhanced items [TEIs]), and more innovative performance tasks to measure writing, research, and problem-solving skills. The tests were first administered operationally in 2015 in a substantial number of states. By the 2015–2016 school year, between the adoption of the consortia tests and states that had implemented their own TBAs, EdTech Strategies (2015) estimated that 85% of the accountability tests in Grades 3 to 8 would be delivered online. Moreover, although many states were using both online and paper-based tests, only three states were not using some form of TBA. Primarily a result of political pressures, state membership in the consortia has waned substantially, with many states since choosing to implement their own TBA programs. Despite the reduction in consortia membership, most states continue to use TBA in whole or in part for their state-mandated accountability assessments (Olson, 2019).

Progressing alongside the development of TBA for state accountability was incorporation into the U.S. national assessment. In 1999, NAEP began a series of studies to facilitate its transition to TBA. NAEP is a congressionally mandated assessment and differs from typical testing programs in that group-level results, rather than scores for individual students, are reported for the nation, states, selected demographic groups, and some large-city school districts. Results are based on samples of schools and students, with a highly efficient matrix-sampling design employed to cover a broad content domain while minimizing testing time for any individual. Although the results do not carry direct consequences for schools or students, NAEP results receive much press coverage and are highly influential in shaping national, state, and big-city education policy.

The NAEP TBA studies investigated the delivery of traditional NAEP paper-based assessments in mathematics and in writing via computer (Bennett et al., 2008; Horkay et al., 2006; Sandene et al., 2005). Computer delivery occurred on desktop computers via the Internet or on disconnected laptops. These studies, which involved administering the assessments in both modes to randomly equivalent samples, examined issues related to comparability of results between computer and paper administration and between laptop and desktop administration, as well as issues related to differential subgroup performance and the role of computer familiarity. Results were largely favorable, suggesting the prospect of transitioning NAEP to TBA while maintaining trend comparisons to previous paper results. A third study explored the use of innovative interactive tasks

to examine problem-solving in technology-rich environments (Bennett et al., 2007, 2010). This study also yielded promising results.

Based on this early work, subsequent research, and developments in technology and testing, as of this writing, the NAEP program has almost completely replaced its paper-based assessments with TBAs. In 2011, NAEP conducted its first operational writing TBA in Grades 8 and 12 (National Center for Education Statistics, 2012). The assessment, which was delivered on NAEP laptops, required students to compose essays in response to prompts, just as in earlier paper administrations. However, because the assessment was based on a new framework, no attempt was made to maintain trends to prior years. In 2014, NAEP conducted the Technology and Engineering Literacy assessment, designed from inception as a TBA (National Center for Educational Statistics, n.d.). This assessment included TEIs as well as scenario-based interactive tasks (see the section on "Innovative Item, Task, and Assessment Types").

NAEP's mathematics and reading assessments at Grades 4 and 8 were transitioned in 2017 and consisted largely of traditional multiple-choice and constructed-response items, along with a limited number of TEIs (Jewsbury et al., 2020). Plans for future assessments in all NAEP subjects presume digital delivery and the increased use of innovative item types. While the current TBAs still employ the standard NAEP matrix-sample designs, adaptive administration has also been explored (Oranje et al., 2014).

In similar fashion to NAEP, such international group-score assessments as PISA and the Trends in International Mathematics and Science Study (TIMSS), are transitioning or have transitioned to TBA. (See Braun & Kirsch, this volume, for a comprehensive discussion of assessments in the international context.) Unlike NAEP, these assessments continue to support both TBA and paper administration because not all participating education systems have the required infrastructure. Consequently, these programs work to maintain comparability of results, both to prior assessment cycles and within a cycle.

PISA, which measures the reading, mathematics, and science skills of 15-year-olds across many nations and jurisdictions, transitioned to TBA for all three content areas (plus financial literacy) in 2015 (Organisation for Economic Co-operation and Development [OECD], 2017). The assessments, which were delivered on school desktops and laptops, consisted largely of TBA analogues to existing paper item types. In 2018, the reading TBA was converted to an MST, whereas the mathematics and science assessments remained linear tests.

TIMSS, which assesses the mathematics and science skills of eighth graders, was also in the process of transitioning to a digitally based version (Mullis, 2019). As of 2019, TIMSS was administered via computers or tablets, as well as in its paper format, with about half of participating education systems utilizing each mode. In addition to traditional items, the eTIMSS digital version included innovative tasks designed to simulate real-world and laboratory situations where students can integrate and apply processes and content knowledge to solve mathematics problems and conduct scientific experiments.

A third international group-score assessment, the Progress in International Reading Literacy Study (PIRLS), has also incorporated a TBA, this one concentrating on the reading skills of fourth graders. In 2016, 16 of the 58 education systems that participated in the paper-based assessment (including the United States), also took ePIRLS, a supplementary measure of online reading skills (Institute for Education Sciences, n.d.; Mullis & Prendergast, 2017). The assessment, which was delivered on either school desktop or laptop computers, consisted of two tasks that required test takers to answer multiple-choice and constructed-response items pertaining to information presented on simulated web pages.

As we look to the future, it seems reasonable to assume that TBAs will move toward a more decentralized delivery model admitting a wider range of assessment devices. As seen with the consortia-developed K–12 tests, beyond requiring a minimum configuration, it is impractical to impose strict control over the delivery device. Within and across U.S. states, schools own a wide variety of equipment. Furthermore, it is not feasible for test sponsors to provide a common device at that scale. Similar considerations exist for international programs like PISA and TIMSS.

Moreover, disruptive events like the COVID-19 pandemic make it evident that consequential testing may need to be carried out remotely at times. In pandemic conditions, gathering in groups is unsafe. Thus, consequential testing may increasingly need to be administered in homes on test taker–owned equipment, where feasible and where the security conditions for such testing are acceptable to both sponsors and users. For the purposes of K–12 assessment, approaches such as remote proctoring involving live monitoring or video capture, as well as recording and analyzing process data (e.g., timing and keystroke sequences), remain controversial as of this writing.

In addition to various exogenous factors, validity and fairness concerns may be moving consequential testing toward a more decentralized delivery model. As more complex constructs are included and tasks are incorporated that require extensive use of and familiarity with the device and interface, issues of comparability and fairness might be better served by allowing test takers to work on their own equipment. In that way, score differences might be more likely to reflect differences in the target constructs rather than also reflecting familiarity with an arbitrarily imposed, standardized device configuration.

To make consequential testing possible in decentralized testing environments, current strategies for ensuring standardization, security, and fairness will need to be adapted. As for K–12 accountability testing, some control over testing conditions may still be imposed through the specification of minimum device characteristics like screen size and resolution, types of keyboards, and transmission bandwidths. Providing free online practice materials and tutorials to familiarize test takers with task types, tools, and interfaces will also continue to be important. But most likely, these strategies will need to be coupled with a careful approach to designing delivery systems, interfaces, tools, and tasks to ensure performance is device agnostic.

Innovative Item, Task, and Assessment Types

In the preceding section, we outlined the development and landscape for consequential TBA. Next, we consider the building blocks of innovation in such assessments—items, tasks, and assessment types.

Among the potential benefits technology brings to assessment design are increased construct fidelity (Russell, 2016), evoking and supporting more complex cognitive processes, and allowing greater observability of the response process. Construct fidelity represents the extent to which the response process evoked by the item accurately reflects the targeted measurement construct. Russell and Moncaleano (2019) identified two components of construct fidelity: the extent to which the problem context is consonant with domain practice and the degree to which the response interaction aligns with the interactions that would occur in a real-world situation. For most constructs, paper-and-pencil items are far from real-world applications in both respects, implying that technology enhancement might allow for increases in construct fidelity.

Complexity of cognitive process relates to the length and depth of reasoning that is required to correctly respond to the task. Criticism has long been made about traditional multiple-choice item formats because they can measure only certain types of cognition and because of the impact that focus may have on teaching and learning (Frederiksen, 1984). Technology-enabled assessments have the potential to measure constructs that involve more complex cognitive processes, such as science practices, collaboration skills, and investigative research (Bennett et al., 2010; Csapo et al., 2012). Through use of simulated problem contexts, computer-based tasks can support an extended problem-solving process more appropriate for these domains from which student cognition might be modeled at a finer grain size (Mislevy & DiCerbo, 2012).

Whereas TBA may be useful in evoking complex cognition, traditional paper-based assessments can also call on higher order response processes through the use of text and diagrams (Bennett, 2012). An advantage of computer-based assessments, however, is the ease with which evidence about response process can be captured. The more interactive an item response, the more process data can be recorded. Thus, technology makes complex response processes more observable. "Observable" does not necessarily mean scorable; as discussed in the section "Response Processes," challenges in interpreting and scoring process data must also be overcome before this evidence can be used meaningfully.

In the following paragraphs, we examine technological innovations in four broad categories: TEIs, extended-interaction items, scenario-based tasks, and simulation-based performance tasks. The first two categories involve innovations at the item level, while the second two encompass innovations for lengthier tasks that may include interdependent components.

TEIs

The most common items used in computer-based assessments are traditional item types with constrained outcome spaces, some of which could have an exact paper-and-pencil replica. These items might take multiple-choice, matching, sorting, or short-answer

formats. They may appear as discrete samplings, independent of one another, or as part of an item set. For assessments offered on both paper and computer, items with exact or highly similar counterparts are desirable to increase the comparability of score meaning between delivery modes. Additionally, these items produce data that fit commonly used scoring and psychometric models, increasing the likelihood that results can be put on the same scale as previous paper administrations.

While the format of such items is basically traditional, technical innovations can be introduced in stimulus presentation or response process, thereby making the item a "technology-enhanced" one. For example, the use of images, audio clips, animation, and video enables a richer presentation of information, which may increase construct fidelity or evoke more complex cognitive processes, even when the response type is highly constrained (Bennett et al., 1999). Figure 9.1 gives an early example that shows how multimedia might be employed in an item stimulus as part of a test of U.S. history knowledge. The item allows the student to view and analyze a primary source in a way closer to that of a historian.

For the response process, the affordances of mouse and touchscreen technology are frequently used to enable drag-and-drop or on-screen drawing as response mechanisms (Zenisky & Sireci, 2002). Drag and drop can be used for questions that ask test takers to move objects into the correct order or for items in which test takers reposition

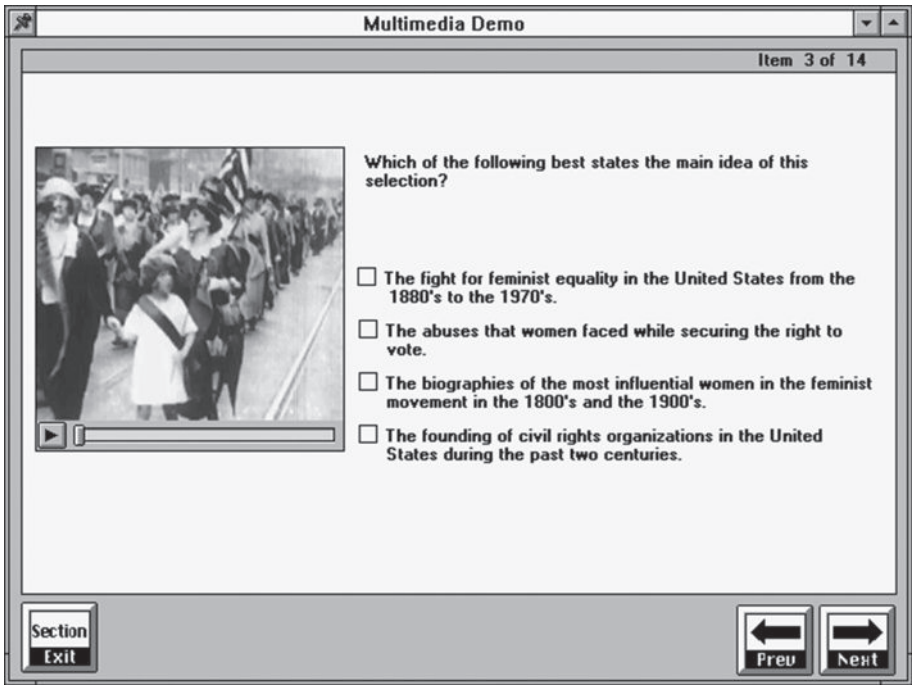

FIGURE 9.1

A Technology Enhanced U.S. History Item Using Multimedia

Note. Copyright ETS© 1999. Reprinted with permission.

objects from one set to match objects in another set. Additionally, drag and drop can access a more continuous response space, such as putting objects onto a map or placing a marker on a spectrum (ATP, 2017; Wan & Henley, 2012).

An example is the PISA 2015 interactive science task, *Fish Farm*. In this item set, students are given a brief context statement about the need to develop a particular type of alternative seafood source and the challenges encountered in operating such a source. Following this context are three independent items (i.e., subsequent items do not build directly on preceding ones or lead to a culminating task satisfying a goal given in the context statement). One of the set's items asks students to help design a sustainable farm ecosystem by dragging organisms into the appropriate tanks (Figure 9.2). The cognitive complexity of this item is fairly high because it "requires students to understand a system and the role of several organisms within that system" (OECD, 2018). Whereas this item could be rendered on paper using a diagram and labels for the organisms, the ability to move the organisms into the fish farm allows test takers to more naturally construct, evaluate, and revise their model.

Russell and Moncaleano (2019) judged that most usage of drag and drop in TBA failed to increase construct fidelity. Interactivity, per se, does not necessarily improve the item. In the fish tank task, the graphical representation allows students to reason

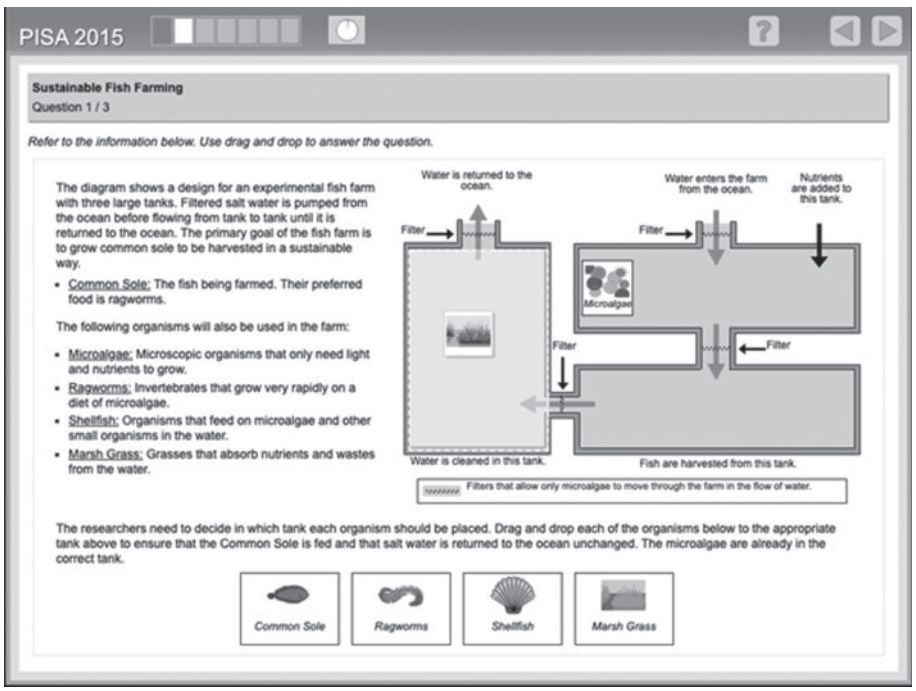

FIGURE 9.2
PISA 2015 Interactive Science Item in Which Students Use Drag-and-Drop Interactions to Complete a Model of a Fish Farm

Note. PISA = Programme for International Student Assessment. From *Try PISA 2015 Test Questions* by OECD, 2018. (https://www.oecd.org/pisa/test/pisa2015/#d.en.537240). CC by 3.0

more directly about the placement and interactions among the organisms. Thus, the drag-and-drop functionality is used effectively to increase the similarity of the evoked cognitive process to ones that a scientist might employ, arguably enhancing construct fidelity.

Extended-Interaction Items

More innovative uses of technology have generally resulted in greater interactivity, providing support for and observation of a more extended response process but still within the context of a discrete, stand-alone item or item set. These items frequently involve constructing a complex response to solve a problem. Test takers might be asked to graphically represent a process, analyze data, or test a hypothesis using a runnable model (a function that accepts input parameters and outputs a result intended to mimic reality, as in the "running in hot weather" example that follows). This type of item is distinguished both by the multiple actions a test taker is expected to take and the depth of cognitive processing associated with those actions.

An example of this class is shown in Figure 9.3. In this PISA 2015 item set, test takers are asked to use a runnable model to answer questions about factors that increase the risk of dehydration or heat stroke for a jogger on a hot day. Test takers are expected to interact with the model, running several trials with different settings before responding

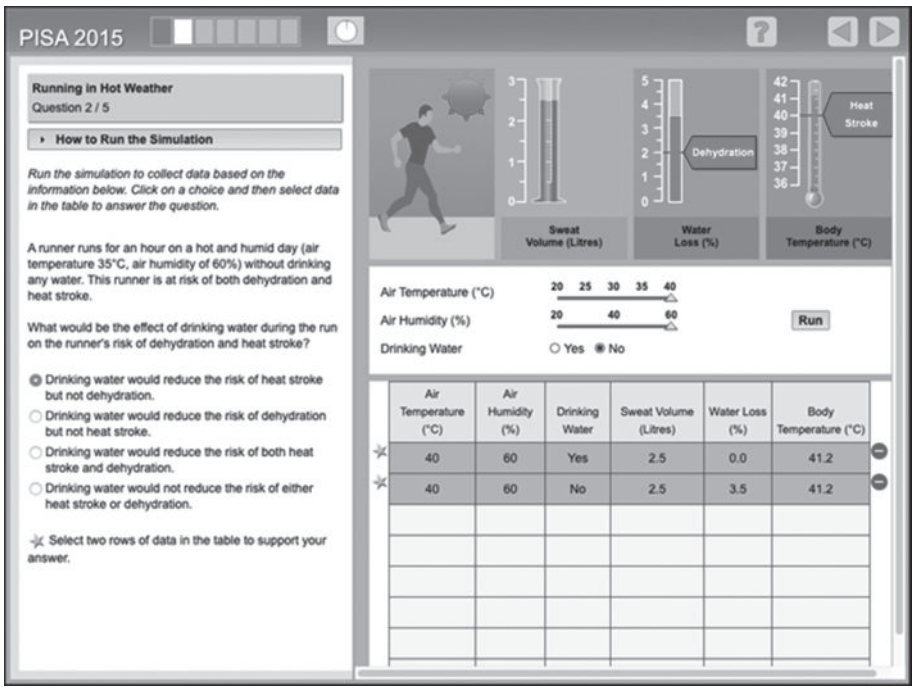

FIGURE 9.3

PISA 2015 Extended-Interaction Item From the *Running in Hot Weather* Set

Note. PISA = Programme for International Student Assessment. From *Try PISA 2015 Test Questions* by OECD, 2018. (https://www.oecd.org/pisa/test/pisa2015/#d.en.537240). CC by 3.0

to the multiple-choice item prompt. The interactions allow students to engage with a real-world problem, while providing evidence of solution process through event logs (discussed in the section "Types of Response Process Data"). Thus, this format permits measurement of a construct that is not possible to assess with a paper test.

Another example of extended interactivity can be found in the PISA 2018 literacy items, which provide a simulated web browser for test takers to employ in researching a topic. While the primary construct is reading comprehension, the item design requires students to follow hyperlinks and use tabs for navigation to relevant content. In this case, the interactivity is intended to increase construct fidelity by replicating the context in which one might undertake information search.

Scenario-Based Tasks

Whereas TEIs and extended-interaction items can provide richer stimuli and the opportunity to observe response processes, their design as independent items (including in sets) limits the depth and complexity of problem-solving. To tap deeper problem-solving processes, assessment designers may turn to scenario-based tasks (SBTs), as defined by Deane et al. (2018), O'Reilly et al. (2019), and others. These tasks are characterized by an overarching narrative, or scenario, which poses a goal for the test taker to achieve. The scenario presentation is followed by a sequence of related technology-enhanced and more traditional items that lead to a culminating performance in which the test taker attempts to satisfy the goal (e.g., a proposal for how a school might use a generous gift, backed by reasons and evidence from given sources). SBTs break the process leading to this goal into steps, each of which is at the same time part of the larger whole but implemented as one or more distinct items.

Some of the tasks included in the NAEP 2014 Technology and Engineering Literacy assessment offer examples. As an instance, the bike lane task consists of five items (see Figure 9.4). Students are first introduced to the motivating problem of safety when riding in a lane adjacent to automobile traffic. Initial items ask students to interact with a runnable road-sharing model that rates bike lane safety based on the manipulable parameters of automobile speed limit and lane width. In later items, students use the information gathered to create safe road designs for cyclists and ultimately to reach the goal of a bike route that optimizes safety, cost, and route length.

The SBT format, when used to measure problem-solving processes, is a compromise between the discrete items of traditional assessment and the open-ended problem-solving characteristic of an extended project. When assessing complex problem-solving, the ability to gather evidence on the process competes with the level of constraint imposed on the problem space. If we allow the student full freedom within the problem space, it is more difficult to detect, understand, and score the problem-solving process used. In unconstrained tasks, the most reliable evidence of competency is typically the outcome. Unfortunately, the outcome may offer little information with which to distinguish and subsequently guide learners within the middle and lower regions of the proficiency distribution. When the task outcome is complex enough to show

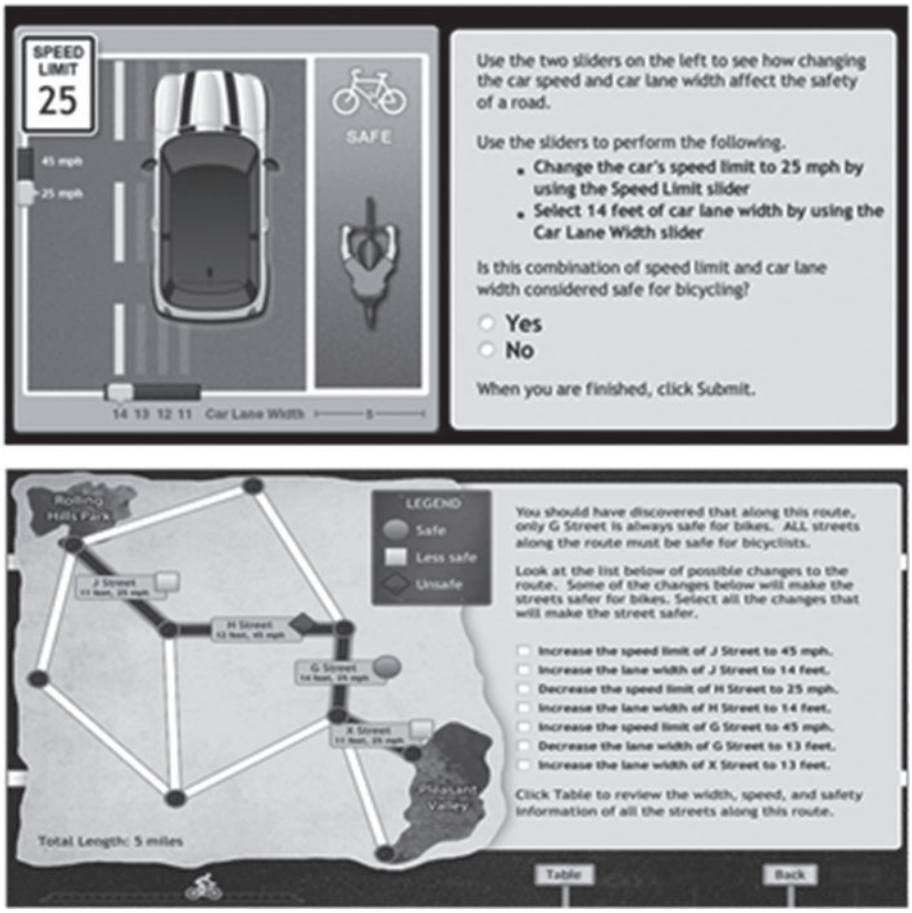

FIGURE 9.4

Selected Screens From the NAEP TEL 2014 Scenario-Based Task *Bike Lane*

Note. NAEP = National Assessment of Educational Progress; TEL = Technology and Engineering Literacy Assessment. From *2014 Technology and Engineering Literary (TEL): Sample Scenario-Based Tasks*, by National Center for Education Statistics, n.d. https://www.nationsreportcard.gov/tel_2014/#tasks/bikelanes

finer gradations of competence, for example, as in an essay or lab report, making such distinctions comes at the cost of greater scoring time and effort.

If we instead discretize the problem-solving steps to allow observation of specific portions of the process, we are inevitably scaffolding problem-solving along a more constrained, and frequently more linear, path. That constraint could improve measurement or, conceivably, undermine it by giving the test taker aids not typically available in the criterion situation. Research suggests that the former situation may be the case. That is, SBTs may in some respects produce better measurement than less-structured performance tasks, particularly in more effectively aligning the processes measured with the intended construct (Guo et al. 2019, 2020; Zhang, Deane, et al., 2019). This result could be due to the initial questions in an SBT helping to activate relevant knowledge and better orient the test taker to the task at hand. In the criterion situation,

similar orientation can come, for example, from consulting with others and viewing examples of quality solutions to comparable problems.

Simulation-Enabled Performance Tasks

While SBTs scaffold the problem-solving process to enable observable evidence about the complex cognition underlying a performance, these tasks are somewhat unnatural. According to Russell's definition of construct fidelity (Russell & Moncaleano, 2019), SBTs may provide a realistic context, but because of their structure, the actions students take might diverge from those that a domain practitioner would employ in their problem-solving.

Performance assessments come closer to actual domain practice, allowing individuals to demonstrate a skill by carrying out a less structured task. A driving test offers a good example. However, although it has high construct fidelity, it requires specialized equipment (i.e., a car), a human examiner, and an individualized administration. A compromise position is offered by TBAs that employ simulation-enabled performance tasks. Such tasks attempt to represent key stimulus features of a given performance situation, calling on competencies that would be employed in that situation, but reduce the resource requirements to more manageable levels.

As an example, the USMLE seeks to assess a potential physician's ability to independently diagnose and treat a variety of patient conditions. As noted, in 1999, the National Board of Medical Examiners incorporated patient case management simulations into the computer-based USMLE (Dillon & Clauser, 2009). In each case management task, the patient's condition changes over simulated time. The test taker can engage in various actions that a doctor might take, including requesting patient history, getting the results of a physical exam, ordering laboratory tests, making a diagnosis, and prescribing treatment. While the "physical exam" is implemented using a set of checkboxes, the lab tests and treatment plan are specified using text input. This format prevents any form of prompting because the test taker will not see a list of available tests or treatment options (unless there are multiple options that match the test taker's input). While these simulations are not similar in fidelity to interacting with a real patient (they do not include any visual representation of the patient or of test results like X-rays), they replicate the main aspects of the problem-solving process used in practice. As mentioned, simulation-enabled performance tasks have also been employed in the architect and accountant examinations, among others (Clauser et al., 2016). (See Margolis et al., this volume, for more on assessment in licensure and certification.)

Less comprehensive simulation-enabled tasks can be found in school testing programs. For example, PISA 2015 assessed collaboration skills using a simulated group assignment in which the test taker negotiates with two automated agents to achieve a specified goal (OECD, 2018). In this instance, all choices made by the test taker were selected responses; however, the choices made changed the task situation.

Of special note in simulation-based performance tasks is that technology plays a key role in observing and making inferences about complex cognitive processes. That is,

the technology environment provides a reasonably realistic problem setting that allows the test taker to work through the problem in a high-fidelity manner. The test taker's step-by-step decisions can be recorded in computer log files, making inferences about their problem-solving process feasible and, in principle, scorable.

Appropriate Use of Innovative Items and Tasks

While innovative item types have important advantages, they are generally expensive to build, more complicated to administer, and less well understood in terms of measurement properties. For consequential testing, such item types should be used only when substantively appropriate, suited to the target population, and logistically feasible.

With respect to substantive appropriateness, as noted, such innovation can improve construct fidelity by enhancing problem presentation or response interaction, thereby evoking more relevant problem-solving processes. Instances include constructs that involve the use of technology itself, such as technical literacy or data analysis, as well as constructs for which the technology can provide a more faithful context for the application of skills, as in scientific inquiry, medical licensure, or remote collaboration (Bennett et al., 2010; Clauser et al., 2016; Csapo et al., 2012).

In addition to fit with the intended construct, appropriate use implies a good fit with the target population. Fit in this context means that all test takers can interact with items in ways that provide evidence of competency. Essential to this premise is an infrastructure capable of validly assessing *all* members of the test-taker population, including those with disabilities or who are English learners (see also Rodriguez & Thurlow, this volume).

Feasibility encompasses assessment design, development, administration, scoring, and reporting. Innovative items can introduce complications in any or all these phases. Design and development time and costs are frequently underestimated by staff members more familiar with traditional assessments. However, the costs tend to be front-loaded because the deployment of innovative item types requires developing or adopting new tools and processes, training assessment developers, designing for accessibility, conducting cognitive labs to evaluate whether the intended processes are evoked, etc. Once this foundation is established, the operational development of innovative items can be streamlined by using appropriate authoring tools and improved through the telemetry (i.e., process data) produced by the items. Finally, scoring and reporting requires appropriate psychometric models to facilitate defensible inferences.

Validity, Modeling, and Analysis Issues

Innovative items and tasks present new challenges for validity, modeling, and analysis, especially for tests used to make consequential decisions. Such innovations demand evidence that new presentation formats and response interactions produce meaningful inferences about the intended constructs. Studies may cover the range of evidence types recommended for validation generally (AERA et al., 2014), but with added attention to whether the response processes evoked are consistent with the target construct

and to what degree construct-irrelevant variance might have been introduced (e.g., Gallagher et al., 2002; Mislevy et al., 1999). For SBTs and simulation-enabled performance tasks, potential unfairness due to Task × Person interaction may also be present (Linn & Burton, 1994; Shavelson et al., 1993).

Task × Person interaction occurs when one test taker is more familiar with, knowledgeable about, or interested in a particular context or topic than another test taker. For shorter tasks, this effect can be easily balanced by having a variety of different contexts or topics across items. SBTs and simulation-enabled performance tasks, however, tend to take significant time, making it impractical to include more than a few tasks. For example, the NAEP Technology and Engineering Literacy assessment tasks take 10–30 minutes each, leaving little room for additional items with contextual or topical variations (a problem NAEP addresses through its matrix design and the reporting of results at the group level, rather than through individual scores).

For TEIs and many extended-interaction items, well-understood psychometric models such as item response theory (IRT) can be used because student response data can be reasonably assumed to meet the statistical assumptions of such models. Any media or interactivity, however, must be carefully evaluated to ensure that their addition does not introduce extraneous cognitive load or other forms of construct-irrelevant variance. Particularly for extended-interaction items, the ease with which the computer interface is learned and employed can significantly affect performance. To ensure fairness, assessment developers must consider the range of experience levels likely in the target population. Similarly, items that use video or audio, or that have fine-motor-skill requirements, must be created to be accessible from the outset. Universal design methodologies are the de facto standard for achieving this goal, including such features as braille, stacked multilingual translations, videos in American Sign Language, and glossaries and test directions in other languages (e.g., Smarter Balanced, 2024).

SBTs and simulation-enabled performance tasks raise significant issues for modeling and statistical inference. While SBTs often contain a set of scored items that resembles a traditional assessment, performance on these items may not necessarily be statistically independent conditional on proficiency. The overarching narrative and topic that define the SBT contribute one source of local dependence that may affect estimates of test precision under standard analysis approaches. When multiple SBTs are used within a larger assessment, a testlet model (Wainer et al., 2000) can be employed to account for the added dependencies among item responses within an SBT.

Another source of dependence might come from how items are related over the scenario. Test takers who draw the wrong conclusions in early items may carry those incorrect ideas forward. Such is particularly the case when the SBT is structured so that later items build on previous results, as in the bike lane task. The effect of previous errors can be mitigated through leveling, in which the test taker is given a correct starting value for the new item. Leveling is controversial, however, because it does not allow test takers to follow their own path, nor does it necessarily eliminate the dependency that could be caused by the test taker's memory of previous reasoning. Models that can

accommodate item dependency, such as Bayesian networks, might be suitable in this context (Almond et al., 2015).

Simulation-enabled performance tasks often have a significant number of correct responses, especially when the response process itself is key to the target construct. In such cases, the log files contain fine-grained details such as mouse clicks, keystrokes, and latencies, which on their own are likely to have very limited meaning. Lending meaning requires thoughtful planning and development, starting with task design (Clauser et al., 2016). Such an approach involves creating a task so that it elicits important features of performance, which can then be aggregated in some fashion for scoring. This approach was used in the USMLE and in the design section of the Architect Registration Examination (Bejar, 1991; Braun et al., 2006; Clauser et al., 2016). New analytic techniques, such as deep learning (LeCun et al., 2015), or novel applications of cognitive-process modeling (LaMar, 2018) may also help in meeting the scoring challenge. The use of response-process data will be discussed further in the section "Response Processes" and automated scoring will be discussed in the section "Automatically Scoring Complex Constructed-Response Tasks."

Comparability of Results and Score Meaning

As noted earlier, many current TBAs can be characterized as first or early second generation—that is, they differ modestly from their paper-based predecessor or current counterparts with respect to item types and the competencies targeted for measurement. However, some of these TBAs also make use of innovative items. Thus, they are beginning to leverage the affordances associated with digital administration. In this section, we deal with the implications of such leveraging for the comparability (i.e., continuity of meaning) of results.

In one use case—national assessment—the comparability of results from new TBAs to prior paper-and-pencil (PBA) results is generally desired. Comparability to PBA versions is sought because one of the defining characteristics of NAEP is the ability to compare the performance of students in the early 21st century—overall and for various subgroups—to cohorts from prior decades. In another common use case, K–12 accountability as implemented through Smarter Balanced and PARCC, the TBAs replace prior, distantly related, paper tests with no expectation of comparability. However, each of these TBAs may coexist with its own PBA counterpart, the choice of examination mode being left primarily to states, districts, and schools.

In both use cases, test takers and test users alike may desire, or expect, that results obtained under either examination mode should be comparable to the other. For group-score assessments like NAEP, recent results obtained with TBA—for the United States as a whole, for demographic subgroups, and for states and districts—need to provide meaningful comparisons to the results from earlier years in which the administration mode was paper and pencil. Changes in scores need to be interpretable because of changes in the target competencies (i.e., to construct-relevant factors), not because of an artifact of the switch from paper to digital delivery. Similarly, for K–12 accountability

tests, differences in results between districts, schools, classrooms, or individual students should reflect differences in the constructs of interest, not factors related to examination mode.

Moreover, in some use cases—for example, TBAs for K–12 accountability—the TBAs themselves allow for considerable variation in administrative conditions. Both the Smarter Balanced and the PARCC tests, for example, are delivered on a range of laptops and tablets, and student responses can be input through keyboards native to the device, an external keyboard, or, in some cases, touchscreens. Despite this variability, test users and takers again expect the results—individual and aggregate—to be comparable across these variations.

The definition of what it means for scores from different versions of a test, or from different tests with similar targeted competencies, to be comparable has received much attention in the measurement literature, initially within the context of more general discussions of score linking and equating (Holland, 2007; Holland & Dorans; 2006; Linn, 1993; Mislevy, 1992; see also Moses, this volume). In this literature, comparability is viewed as a matter of degree, with the strongest level being achieved when scores from different tests can be considered *interchangeable*. Less-stringent degrees are achieved through other forms of scale alignment such as concordances, statistical relationships between the results of assessments of related constructs that hold in specific populations.

Interchangeable scores can often be obtained in the context of consequential tests, like SAT and ACT, which traditionally produced different parallel forms administered in the same mode (primarily PBA) for use at each scheduled test administration. Such forms are constructed so that the assessed content and psychometric characteristics are tightly controlled to enhance comparability. As a result, the test takers at a given level of proficiency can be expected, on average, to achieve the same test score and to be measured with the same degree of precision, regardless of the test form administered. Moreover, score meaning—in terms of the competencies measured—can also be assumed to be the same, regardless of the form given.

It is noteworthy that even within this highly constrained context, it is rarely the case that the *raw scores* from different forms (e.g., simple number-correct scores or even IRT-based estimates) can be treated as interchangeable. However, with proper data collection designs and appropriate analysis procedures, scores from these alternate forms can be *equated* (i.e., adjusted for any unintended differences in difficulty and/or expressed on form-invariant scales, like the well-known ACT and SAT scales). For most practical purposes, after equating, the resulting scale scores from different forms of the same test can be treated as interchangeable in terms of the constructs measured, as well as psychometric characteristics, when making inferences about individual test takers or groups.

In considering the comparability of scores from PBA and TBA forms, interchangeability may be a reasonable and desirable goal when the content, item types, and delivery paradigm (e.g., linear versus adaptive) of the two modes are kept as close as possible. In situations

where the identical (or nearly equivalent) set of items is being delivered, there is a long history and considerable body of empirical evidence with different kinds of tests and testing populations, often referred to collectively as *mode comparability studies*. Though findings from this body of research are somewhat mixed, the evidence suggests that—as is often the case with different PBA forms of the same test—some degree of equating adjustment to TBA raw scores may be required if results are to be treated interchangeably with those of the corresponding PBAs. Papers by Paek (2005), Kingston (2009), Drasgow et al. (2006), Lottridge et al. (2010), and W. D. Way et al. (2006) offered pertinent reviews for tests that provide individual scores. Jewsbury et al. (2020) gave an additional recent example from NAEP pertinent to group-score assessments.

In some use cases (e.g., the Smarter Balanced K–12 accountability tests and the PISA transition to TBA), design differences between PBA and TBA also exist—for example, the paper version being linear and the computer one item-level or multistage adaptive. Changes in test delivery that involve an item-level CAT or an MST could, in some cases, be expected to affect the psychometric properties of scores, usually by improving the precision of results for low- and high-performing test takers relative to test takers in the middle of the score distribution. Hence, strictly speaking, scores from the linear PBA and adaptive TBA may not be interchangeable in the same sense as scores on parallel forms from the PBA test, the differences in precision at particular ability levels being intentional and desirable. However, through the imposition of content and format constraints and the conduct of appropriately designed equating studies (see, e.g., Dorans, 2000; Schaeffer et al., 1995, 1998), adaptive versions of linear PBAs can be instantiated such that scores can be treated as equivalent measures of the same construct for practical purposes.

As of this writing, however, the TBA versions of a test can be expected to differ at least to some degree from their PBA counterparts in ways that go beyond a simple move to adaptive delivery. Even first- and second-generation TBAs frequently contain some number of TEIs for which exact PBA counterparts do not exist. As a result, subtle cross-mode differences in the assessed target competencies are inevitable. Similarly, construct-irrelevant factors—such as familiarity with and ease of working within the digital and paper testing environments, respectively—might also reduce the degree of comparability. Different approaches to scoring constructed-response items represent an additional threat to interchangeable scores. For example, absent training, human raters have been found to differentially grade handwritten versus computer-entered responses (Russell & Tao, 2004a, 2004b; Sandene et al. 2005, pp. 14–15).

The *Standards* (AERA et al., 2014, p. 105) requires a clear rationale and supporting evidence for claims that scale scores earned on alternate forms of a test may be used interchangeably. Given the goal of leveraging the affordances of digital delivery, achieving interchangeability is unlikely, perhaps unnecessary in many use cases, and probably undesirable if it limits improvements in measurement precision and better representation of important constructs. However, even when strict equivalence is not claimed, the *Standards* requires a rationale and direct evidence of the *degree* of

score comparability commensurate with intended uses and claims (p. 106). Similarly, U.S. Department of Education guidance on peer review of state assessment systems (U.S. Department of Education, 2018) suggests that when a state administers different versions of its assessment (e.g., TBA and PBA, or TBA on different devices), the state should provide comparability evidence generally consistent with the expectations of current professional standards.

What sorts of evidence traditionally have been, and should be, provided to document the degree of comparability between TBA and PBA results? Whereas the *Standards* contains no specific guidance, there is a substantial literature that addresses the topic generally. Some of this literature is located within the broader context of establishing comparability of scores obtained under standardized conditions to those obtained from test variations such as accommodations, adaptations, or different languages, where interchangeable scores are rarely practically achievable. Particularly pertinent discussions can be found in Kolen (1999), Wang and Kolen (2001), Sireci (2005), Winter (2010), Lottridge et al. (2010), Randall et al. (2012), DePascale et al. (2016), and Berman et al. (2020).

There is considerable agreement in this literature regarding the sources of evidence required. These sources include studies that examine:

- similarity of the dimensional structures;
- similarity of item-level psychometric properties (e.g., classical test theory indices of difficulty and discrimination, IRT curves, IRT model fit in cases where the TBA and PBA versions contain identical or corresponding items);
- similarity of predictive and concurrent statistical relationships between assessment scores and other related educational variables;
- similarity of measurement precision through comparisons of overall standard errors of measurement and standard error curves across the proficiency range;
- similarity of overall score distributions, including means, degree of dispersion, and shape; and
- similarity of score differences and extent of differential item functioning for important subgroups defined by gender, race/ethnicity, disability, socioeconomic status, or computer familiarity.

Whereas a comprehensive evaluation of comparability would involve all these sources of evidence listed, rarely in practice is that possible. In analogous fashion with validity, degree of comparability remains an integrative judgment by test professionals, test users, and other stakeholders as to whether claims based on comparisons of PBA and TBA versions are adequately supported.

From a fairness perspective, perhaps the most important sources of evidence for tests used for consequential purposes are associated with what Winter (2010) labeled "score-level comparability" (the last three sources listed above). Direct evidence is usually gathered from mode comparability studies in which TBA and PBA results are obtained from groups that can be assumed to be equivalent with respect to the

distribution of the target competencies associated with the test. Data from such studies serve two roles. The first role is to provide evidence of score-level comparability. The second role is to, when necessary, determine an adjustment from one or the other modes to account for the effects—that is, differences in overall test performance (means, standard deviations, or shapes). Depending on other sources of evidence, the adjustment can, in the best case, render the scores effectively interchangeable or, where claims of interchangeability are not justified, increase the degree of comparability and fairness inherent in score use.

Mode-Study Data Collection Designs

Mode comparability studies have made extensive use of the data collection designs traditionally employed for equating and linking (see, for example, Kolen & Brennan, 2004; and Moses, this volume). *Single-group designs* have many advantages, in principle. In these designs, test takers are given both the PBA and the TBA versions, usually with the order of administration randomly counterbalanced (e.g., Gallagher et al., 2002). Because the same test takers are tested in both modes, issues related to sampling variability between groups are minimized. Moreover, this design provides the most direct evidence of the construct equivalence of the TBA and PBA versions. The correlation between scores in the two modes and the similarity of internal dimensional structure, psychometric properties, and relationships with other variables can all be directly evaluated *on the same group of test takers*. If IRT methods are employed, the appropriateness of jointly scaling both tests can also be directly appraised by examining model fit.

However, there are several challenges with implementing such designs. It is often not feasible to administer tests in both modes to the same test takers because of concerns about burden and so-called *order effects* (i.e., the fact that the relationships between the TBA and PBA versions may differ depending on administration sequence). Lottridge et al. (2010) described several such studies where differential order effects were found.

A somewhat more practical alternative that offers many of the advantages of single-group designs is the *random-groups design* (i.e., where test takers are randomly assigned to administration mode). Because test takers are assessed in only one mode, issues of burden and order are not relevant. When data from large, randomly equivalent groups of test takers are available, almost all the sources of evidence recommended above can be produced, the major exception being correlations and joint IRT scaling. NAEP has relied, and continues to rely, on such designs in carrying out its transition from PBA to TBA (e.g., Bennett et al., 2008; Jewsbury et al., 2020). As part of its 2015 field test, PISA also employed this design to aid in its TBA transition (von Davier et al., 2019).

Operational research from NAEP, described in detail by Jewsbury et al. (2020), is a particularly good example. Data from the random-groups design were used to adjust for mode effects associated with TBA, as well as to provide comprehensive analyses (including classical test theory and IRT), documenting the evidence for the comparability of NAEP results across modes after adjustment. In this study, NAEP administered

the PBA and TBA versions of its fourth- and eighth-grade 2017 mathematics and reading assessments to random samples of test takers, with random assignment to mode carried out largely within each NAEP-participating school. This *within-school* design was chosen to produce highly similar samples, taking each mode at all the jurisdiction levels for which NAEP reports results (nationally, by state, and for large urban districts), and to give reasonable statistical power for detecting mode differences at the state and subgroup levels.

National-level results comparing item-level statistics across modes (biserial correlations, differential item functioning statistics, IRT parameter estimates), as well as various other psychometric characteristics (e.g., dimensionality), provided solid evidence that the two versions were measuring highly similar target competencies. However, other test-level analyses indicated that, without adjustment, results from the TBA versions of the 2017 reading and mathematics assessments would be systematically lower for both fourth- and eighth-grade test takers, with larger mode effects for fourth graders. After making a *single adjustment* to the TBA NAEP scale-score distributions to equate the mean and standard deviation for the national samples taking the assessment in each mode, results were shown to be highly similar for the major national reporting subgroups (e.g., sex, race/ethnicity), for states, and for participating large urban districts. That is, with very few exceptions, the observed differences due to mode were within the bounds of sampling error. The few statistically significant differences that were observed showed little consistency across grades and subject.

PISA employed a somewhat different approach (von Davier et al., 2019). In the 2015 field test, a number of countries provided data from randomly equivalent samples of 15-year-olds for both PBAs and TBAs of reading, mathematics, and science. The investigators combined descriptive analyses—that is, visual inspection of residual plots and indices of item-level model fit—with statistical evaluation of a series of constrained IRT models that they refer to as "mode-effect models." The mode-effect models imposed increasing degrees of measurement invariance. These analyses identified some items for which comparable functioning could not be supported. However, for most items, the evidence suggested that the IRT parameters were equivalent across modes. These items were then treated as equivalent in the operational 2015 analyses so that results from both PBA and TBA could be reported on a common scale comparable to prior PISA assessment cycles (OECD, 2017, chap. 9).

Whereas the random-groups approach can be quite effective, it too may frequently be intractable because it relies heavily on random assignment and requires large sample sizes for sufficient statistical power. Whereas NAEP was able to achieve both requirements, those requirements may not be easily met in other operational contexts (e.g., where the test-taking population has uneven access to computers or where imposing an administration mode is not acceptable to the test taker or institution). In such situations, data collection by necessity will involve self-selected, *nonequivalent groups*, that is, groups in which similarity of the distribution of target competencies for the TBA and PBA versions cannot be assumed.

With nonequivalent groups, mode comparability studies have typically adopted one of two strategies (or their combination). The first strategy is to use statistical approaches that employ demographic and other related performance data to control for preexisting group differences in the targeted constructs. Approaches such as propensity-score matching (Rosenbaum & Rubin, 1985), coarsened exact matching (Iacus et al., 2011), or weighting-based approaches (Haberman, 2015) can be employed to create *pseudo-equivalent groups.* These matched groups can then be used to conduct the desired comparability analyses, including the calculation of any mode adjustment. Numerous examples of this approach can be found (Lottridge et al., 2010; W. D. Way et al., 2006, 2007, 2008; Yu et al., 2004).

A second strategy is to rely on *common-item assumptions*—that is, the assumption that identical or nearly identical items appearing in both modes exhibit the same psychometric characteristics (e.g., difficulty and discrimination). When employed in the context of nonequivalent groups, this strategy can be thought of as a variant of the *common-item* or *anchor test nonequivalent groups design* (Kolen, 2007). Numerous IRT methods can be applied, including concurrent calibration and approaches based on separate calibration of PBA and TBA data, followed by some form of parameter linking on test characteristic curve transformations. In addition, non-IRT methods can be brought to bear (see, for example, Kolen & Brennan, 2006, chaps. 4–6). Usually, the analyses proceed by first assuming that all identical/similar items function comparably and then selectively relaxing this assumption based on various diagnostic indices, such as model fit and analysis of outliers.

As a basis for evaluating comparability, both approaches to the nonequivalent groups situation come with challenges that can compromise their effectiveness. Statistical matching requires sufficiently strong ancillary data related to test performance that accounts for the differences between the groups with respect to the targeted competencies. When such ancillary data are not available, the matching will not produce equivalent groups. As such, conclusions regarding comparability, or adjustments to achieve comparability based on equating the score distributions in the matched groups, may be suspect. It is important to note that ineffective matching could potentially affect conclusions in both directions. That is, one could be led to conclude that scores are not comparable when in fact they are or to miss a systematic directional bias that was inadvertently removed through faulty matching.

A possible example of this situation can be seen with the PARCC K–12 reading and mathematics accountability tests, which existed in both TBA and PBA form. Though not strictly identical with respect to all item types, the two modes were intended to produce comparable results. Comparability studies were conducted based on field test data in 2014 (Brown et al., 2015) and on the first year of operational data in 2015 (Liu et al., 2016). The first of these studies was originally intended to use a random-groups design, but challenges in implementation made it necessary to rely on a post hoc matching approach. In contrast, the second study was designed to employ nonequivalent groups with propensity-score matching. In each study, a comprehensive analysis was done

using most of the generally recommended evidence types (AERA et al., 2014). Both studies presented solid evidence that highly similar constructs were being assessed by the two modes. Moreover, though small mean differences usually favoring PBA were noted for particular test subject/grade combinations, the differences in average scores between modes were, for the most part, negligible. Thus, the general conclusion was that results could be compared across modes for accountability purposes.

Despite this conclusion, subsequent studies and analyses of operational PARCC results have suggested that, in at least some districts and states, scores on the PBA version were systematically higher to a greater degree than would have been expected from the mode comparability study results (Backes & Cowen, 2018; Duque, 2017). One reason that evidence of noncomparability has surfaced may be that the earlier matching procedures did not work as intended, a possibility expressed in the 2016 PARCC Technical Report (Pearson, 2017, pp. 143–144). However, other considerations—such as differences in the relationship between TBA and PBA results across groups, states, and time—represent alternative, or at least contributory, factors. Several of these considerations are discussed in the following paragraphs.

Relying on common-item assumptions effectively assumes that, at least on average across all the items common to both modes of presentation, consistent *main effects* on item difficulty and discrimination favoring one or the other mode are negligible. Studies employing this approach are most typically carried out within an IRT framework. In one variation, IRT item parameter estimates based on data from one mode (the reference mode) are applied to the data from the other mode. Analyses of model fit are done to identify items for which the assumption of common parameters appears untenable, and these items are then deleted from the set. For these identified items, as well as items unique to the other mode, separate item parameter estimates based on data from the new mode are obtained, with the parameters for the remainder of the items fixed at their reference-mode values. The full set of item parameters is then used to produce results that are directly on the reference-mode scale.

In situations where the tests have been carefully created to measure similar constructs across modes and the assumptions of negligible main effect are tenable, this approach can achieve comparability of score distributions, as well as provide clues regarding item features that interact to introduce cross-mode noncomparability (e.g., when the mechanics of response entry for a TEI are more complicated than answering the analogous question on paper). The latter information can be quite valuable in producing future comparable versions. However, the assumption of negligible main effects cannot be effectively evaluated based on the mode-study data collected with the nonequivalent groups design. This situation stems from the fact that main effects on item parameters due to mode are not separable in the IRT analysis from group differences on the target construct since the two causes are perfectly confounded. Moving forward, these differences will manifest themselves as a systematic source of noncomparability between scores in the two modes. Attempts are sometimes made to combine the common-item approach with matching strategies to disentangle the mode effects from

group differences (ETS et al., 2016; Liu et al., 2016). However, the effectiveness of this strategy is subject to the limitations discussed previously.

Much of the conceptual framework and methodologies within which psychometricians have approached issues of comparability (particularly of score distributions) was developed when consequential testing was done as PBA under generally strict standardization of administration conditions and testing formats. To be sure, testing conditions across PBA sessions varied in small, presumably inconsequential ways. The basic PBA administration infrastructure could be assumed to be stable, reasonably homogeneous, and familiar to the vast majority of current and future test takers. These assumptions of homogeneity and stability undergirded the argument that standardized test scores were fair, valid, and reliable and that results from such assessments could be confidently and meaningfully compared over time.

Early forays into consequential TBA (e.g., for military selection, licensure, graduate admissions) occurred largely with adults at test centers where the variability in equipment (desktop computers) was limited and under the control of the testing organizations. Given these conditions, it was also reasonable to assume that the TBA administration infrastructure was stable and homogeneous. Thus, similar claims for the fairness, validity, and comparability of TBA results could be made. Against this backdrop, historical approaches to evaluating comparability (particularly of score distributions) seemed appropriate. Unintended differences in difficulty due to mode, for example, could be identified and adjusted for *once*, much the way adjustments have been made to alternate forms of admissions tests like the SAT and ACT. That adjustment could then be applied to results produced by future forms of the TBA, rendering them comparable to past PBA results and to one another.

However, the technology landscape for TBA is different and far more variable, particularly in K–12, where laptop computers and tablets are used for instruction as well as testing. Differences in displays (size, resolution), input mechanisms (keyboards, mouse, touchscreen), and operating systems carry the potential to inadvertently introduce construct-irrelevant differences and impact psychometric properties, not only in comparison to PBA versions, but also among the "same" TBA taken on different devices.

NAEP provides a good example of the challenges that consequential testing programs face. To remain relevant, NAEP has already transitioned many of its assessments (reading, mathematics, U.S. history, civics, geography, and science). Up through 2024, NAEP's operational approach has been to standardize by administering the assessment on the *same* device configuration brought into the participating schools by NAEP personnel. This approach is an attempt to provide, at any given time, a stable and consistent infrastructure capable of delivering the full range of NAEP assessment tasks and of producing comparable scores.

However, maintaining a stable delivery architecture *over time* is becoming untenable. Hardware, software, and interface design life cycles make change inevitable. New delivery devices, operating systems, and assessment software are periodically

introduced. Concomitantly, as NAEP gains experience with its digital delivery system, the program itself will want to make user interface and task design improvements. These changes will affect to varying degrees how the test taker interacts with the test, which could in turn materially change the meaning of results over time. This evolution will be continuous and inexorable.

Of course, when NAEP was a paper-and-pencil assessment, changes in assessment procedures and content were also periodically introduced. At these limited change points, *bridge studies* (using random-groups designs) were conducted to adjust for the potential impact of the changes on assessment results and to confirm the valid reporting of trend. For PBA administrations, results from the sequence of assessments occurring before and after the change could be safely assumed to be comparable since the basic underlying test delivery system (e.g., booklets, pencils) represented a constant delivery infrastructure. However, in the TBA era, conducting mode-comparability studies at major change points may not be practical, economically feasible, or enough to confirm valid trend reporting. Even with a relatively stable assessment in content and procedures, the constant evolution of the delivery infrastructure may be too great.

In contrast to NAEP, Smarter Balanced and PARCC have taken a different approach, electing to support TBA delivery across a wide range of digital devices. This strategy presents a different challenge: how to accumulate and present evidence of comparability across a potentially large number of different devices. As is true for mode effects, the research on device effects has generally used some version of the designs described earlier. Reviews by DePascale et al. (2016) and W. D. Way et al. (2016) summarized some of the key findings about the impact of such factors as screen size, input device, and item type, as well as how these factors interact with such things as content area. But given the continuing rapid evolution of TBA, a strategy of amassing comparability evidence (as historically done for PBA and TBA) would appear to be Sisyphean considering the plethora of existing TBA device configurations. And as noted, the COVID-19 pandemic brought the administration of some TBA programs to the home, where an even greater array of device configurations may be found.

Considering this reality, a more expansive approach to comparability is required. W. D. Way et al. (2016) argued that we have moved from an era of standardization to one of personalization. In that latter era, consequential testing is more accurately viewed as a collection of variations, many of which are intended to adapt the experience to the individual in a way that maximizes their access to assessment content (see also Bennett, 1999, regarding "generalized accommodation," and Bennett, 2024). From this point of view, it will be necessary to supplement, if not entirely replace, studies documenting the comparability of results between variations with mixed methods interdisciplinary approaches. Such approaches should provide test, item, and interface design principles usable across a wide range of assessment situations. The knowledge obtained would inform the engineering of personalized testing conditions in a way that maximizes validity and fairness by increasing the likelihood that results can be treated as comparable across a range of devices.

As W. D. Way et al. (2016) so aptly put it,

> What is clear is that the evolution of technology will only continue and that schools will continue to adopt different technologies in different time frames and will have little patience for a measurement field that is unprepared to accept these technologies for testing purposes. Establishing a framework and a process for evaluating new devices and new technologies is perhaps more important than understanding the impact to comparability of any specific device or technology. (p. 274)

D. Way and Strain-Seymour (2021) have taken a step in that direction by proposing a framework for device and interface features that might affect test performance and, thereby, perhaps offer a path to greater comparability through the personalization in the manner W. D. Way et al. (2016) suggested. That idea is consistent with an emerging alternative view of standardization in educational testing more generally. For example, Sireci (2020) argued for UNDERSTANDization, in which the first step is to appreciate the implications of diversity within the student population and then critically evaluate the ways in which traditional standardized procedures may lead to biased estimates for some individuals and groups. The third step involves adjusting those procedures to eliminate potential biases.

Some precedent for this view exists in the way in which testing programs now handle accessibility. For example, Smarter Balanced offers accessibility features in three categories: accommodations (available to those students with documented need through either an Individualized Education Program or a 504-accommodation plan), designated supports (available to any student for whom school officials have indicated the need), and universal tools (available to all students; Smarter Balanced, 2023). Universal tools include English glossary (i.e., pop-up definitions for selected construct-irrelevant words), highlighter, strikethrough, zoom, and notepad. Among the designated supports are masking, color contrast, text to speech (for all items except reading passages), glossary translation in 10 languages and several dialects, and translated test directions in 19 languages. Accommodations are provided through braille, closed captioning (for listening items), American Sign Language video presentation (for listening and math items), and text to speech (for reading passages in Grades 6 and above), among other mechanisms.

Smarter Balanced test administrations, thus, may vary considerably across students depending on the category of accessibility features for which they are eligible and the features utilized within that category. This variation is intended to follow the guiding principle that, when selectively metered, accessibility supports can contribute to more valid measurement because the assessment is more appropriately customized to student characteristics.

The beginnings of a scientific framework within which to conceptualize comparability for students with special needs may exist in the research on the so-called interaction hypothesis and on differential boost. The interaction hypothesis indicates that accommodated tests (e.g., the availability of text to speech) should result in students

with special needs achieving higher scores than they otherwise would, coupled with no difference for general education students taking the same accommodated test. In contrast, differential boost focuses on the expected increase in scores for students legitimately needing the accommodation, ignoring any effect (or lack thereof) on general education students. Sireci and O'Riordan (2020) provided a comprehensive review of issues related to accommodations and comparability, Sireci et al. (2005) gave a discussion of the interaction hypothesis among students with special needs, and Pennock-Roman and Rivera (2011) did the same for English learners.

Though comparability is emphasized in many consequential use cases, it is important to acknowledge that too large an emphasis may have the unintended negative consequence of inhibiting advances in measurement science and practice. Comparability of scores to previous versions of an assessment may be undesirable if it is obtained at the cost of innovation in test design and task types that provide the potential for improved measurement precision and tapping important constructs. Thus, as testing programs transition from PBA to TBA, they must carefully weigh the costs and benefits of maintaining comparability to past results.

When severing comparability to past results, new reporting scales are typically introduced, which can be extremely disruptive to test takers, test users, and other stakeholders. Testing programs that do so often provide concordance tables showing pairs of scores, new and old, having the same percentile rank in a particular population of test takers. Such statistical relationships can be helpful during the transition when stakeholders may be required to make consequential decisions about test takers, some of whom have scores on the prior reporting scale and others on the new scale. Concordances, however, carry their own risks of misuse and misinterpretation. The reported statistical relationship does not imply interchangeability of results and score meaning. In situations where the knowledge, skills, and abilities measured by the old and new version of the test are substantially different, the statistical relationships may not hold in populations different from those on which the concordance was established or in subgroups of the population. (See, for example, Dorans & Walker, 2007; Pommerich, 2007; and Sawyer, 2007, for a discussion of the uses and limitations of concordances.)

Strategies like instituting innovations incrementally to avoid disrupting comparability at a single time point represent a different approach to balancing the trade-offs. The strategy being used by NAEP for its assessments in reading, mathematics, and science appears to follow such an evolutionary approach. In its initial stages, the transition concentrated on moving the current paper assessment to digital delivery. Later stages introduced innovative item types like SBTs and simulation-enhanced performance tasks. Subsequent stages include a 2025 field test of web-based delivery on school-provided devices and possibly further explorations of adaptive testing.

Introducing such innovations in stages—and evaluating their impact on the comparability of results through careful experimental study and other empirical approaches (see, for example, Jewsbury et al., 2020)—will help NAEP develop validity evidence for supporting construct-based inferences from existing trend lines. If supported

by empirical study, changes in results from the original assessments to the newer innovative ones can be interpreted as due largely to construct-relevant factors, overall and for most or all subgroups of interest. The evidence supporting construct-based inferences is likely most sound for those assessments close to the transition points—and to the comparability studies—where the innovations were introduced. The validity of construct-based inferences may be less well supported for assessments farther apart in time, given the potential cumulative effect on score meaning of the multiple innovations and the absence of studies directly comparing the modified and original versions. Given the rapid changes in technology and the potential advances in assessment practice, increased ambiguity as to the meaning of changes in results over longer time spans may be the price paid for maintaining a relevant assessment system.

In other instances, where an evolutionary approach is not practical or desirable, breaking with the past may represent the wiser course. NAEP again provides an example where decisions to forego comparability with prior assessments were made. The most recent instance concerns the transition of the writing assessments at Grades 8 and 12 from paper to TBA in 2011 (U.S. Department of Education, 2012). The new NAEP writing assessment was based on a different framework than that used in prior years (1998, 2002, and 2007). The implementation of that framework required the assessment to reflect the reality that most writing in the second decade of the 21st century in educational and work settings was already done with the aid of technology. TBA delivery included the concomitant introduction of resources such as a thesaurus, as well as such common computer tools as spell check, cut, copy, and paste—all of which had no analogue in the NAEP PBA context. In addition, the new framework called for changes to the assessment tasks that required a purpose for writing and a specific audience to be addressed. NAEP decided that the changes reflected important social and educational developments and chose to begin new trend lines without empirical study of comparability. The loss of comparability to prior assessment results was seen as a necessary and desirable trade-off in return for maintaining the program's relevance and position as an innovation leader.

Response Processes

Whereas a loss or reduction in the comparability of results and score meaning to PBA might be viewed as a negative effect of the transition to TBA, a more positive outcome relates to the availability of data about the response process. The response process describes the cognition in which a test taker engages when encountering an item stimulus, formulating an answer, and entering that response (Wilson, 2005; Ercikan & Solano-Flores, this volume). For some constructs, the test taker's cognitive process is itself a relevant target of measurement, either instead of, or in addition to, the final product of that process (e.g., in medical patient management, scientific inquiry). This section discusses how technology enables the observation and analysis of evidence concerning response processes. Such evidence can be captured and analyzed from the data collected in any type of TBA, even ones that include only traditional

multiple-choice items. However, the potential value, as well as the challenges of such capture and analysis, increases with task complexity.

The simplest response processes might involve multiple-choice tests of declarative memory: The respondent is asked to recall a fact, find the matching response from a fixed list, and mark the appropriate option. The response process to an algebra item might be more complex, perhaps involving multiple steps of problem translation, integration, solution planning, and execution. A yet more complex response process is evoked when a student is asked to write a persuasive essay. This task might bring into play such aspects as choosing the thesis, outlining and planning the argument, drafting text, and editing and revising that text. A final type of response process is exemplified by a simulation-enabled science performance task. Here, selecting a hypothesis, choosing variables, designing and running an experiment, and interpreting the results would all be relevant.

Next we describe the diverse uses of response process data, after a short introduction to the types of data that TBAs can collect. These uses include validation, quality control, security, and new insights into group and individual performance.

Types of Response Process Data

TBAs can, in principle, provide a stream of process data to supplement the traditional scored response. Log files are commonly generated, recording key events in the response, as well as in the navigation of the assessment. Design of good log files is a new but important part of assessment development (Hao & Mislevy, 2018). On the one hand, whereas every keystroke, finger swipe, mouse action, and latency can be recorded, the resulting data will constitute a large, potentially uninterpretable, collection. Recording too little, on the other hand, may lose valuable information. In general, it is recommended to record as much information as possible because new analytic techniques are emerging for dealing with the data.

No matter the actual item format, TBAs can provide latency data for each student on each item. For example, when a student loads an item, when they select a response, and when they submit can be recorded. For reading comprehension items, timestamps can denote how long a test taker stayed on a single screen; for a writing task they can indicate how long the test taker paused before beginning to write and between characters, words, and sentences. The more interactive an item, the more timestamped events can be collected.

As the called-for interaction increases, more details of the process used to complete a task can be inferred from the log file. In writing tasks, for example, the recorded keystrokes can allow for a real-time recreation of exactly how an essay was composed. Mouse actions can also be recorded for items that involve more complex response formats. For highly interactive tasks, such as simulation-enabled performances, each step can be logged, providing a fine-grained record of the test taker's path to solution.

It is useful to consider the different levels of inference that might be made about a test taker's response process from log files. The lowest level is descriptive, based on the raw

data. These data might include each keystroke, cursor movement, and mouse action, along with assessment-generated events such as the running of a video clip. While data at this level are often more difficult to interpret, they may be used to create a playback, or description, of a student performance. Such playbacks can be used for quality assurance, validity studies, and performance ratings by humans and as a formative resource for instructors or test takers. This level of data is also objective. If the log records a mouse click on a specific button at a particular time, it is almost certain that the click occurred as recorded.

A next level of inference is the action or decision layer. At this more semantic level, mouse clicks or keystrokes may be interpreted as indications that the test taker decided to take an action within the task. Thus, a mouse click becomes "selecting response X" or, in the context of a science simulation, "running a new trial." While these inferences are relatively low level, one cannot be sure that the test taker intended the interaction as it might be interpreted.

The highest semantic level identifies strategies, plans, and knowledge. This level requires the combination of multiple events into a meaningful pattern in the context of the task. For example, we might infer that a student implemented a "control-of-variables" strategy to test a particular hypothesis (LaMar et al., 2017) or that a student wrote an outline as a plan for their intended essay.

It is worth noting that timing data can strengthen or weaken inferences. For example, we might infer that a test taker did not take the time required to develop a strategy or even read the question prompt. This judgment would indicate that the events recorded in the log file were more likely associated with random noise than with planful actions and strategies.

Utility of Response Process Data

Because the test-taker response process is central to the integrity of educational measurement, process data have multiple uses throughout the assessment life cycle. Assessment designers have long employed student cognitive labs to test and refine items in development (e.g., see Connolly & Wantman, 1964, for an early example). Think-aloud protocols or retrospective interviews aim to make observable how the test taker works through questions so that item functioning can be improved.

With the help of automatically generated process data, far larger test-taker samples can be evaluated than are possible with these more labor-intensive methods. In the early stages of assessment design, such data can be used to identify usability problems with the computer interface and to select which tasks are most likely to evoke response processes involving the target construct. Items that allow for shortcut solutions or the use of construct-irrelevant skills can be identified and modified or dropped. Whereas gathering some types of validity evidence is a routine part of assessment design (e.g., item alignment, cognitive labs), most evidence is generally gathered from pilot, field test, and operational administrations because those events provide the sample sizes needed for psychometric analysis. Examining automatically gathered process data could

complement these analyses (Ercikan & Pellegrino, 2017). For example, Zhang, Deane, et al. (2019) compared the psychometric properties of a scenario-based English language arts assessment design to alternative structures and found that a single scenario functioned best with the essay appearing after (rather than before) a series of lead-in tasks. These psychometric analyses were complemented by process investigations that evaluated *how* students wrote their essays. Those investigations showed that the scenario-based design reduced the impact of general writing fluency processes on essay score, thereby presumably giving students with less keyboard facility or lower verbal fluency more opportunity to display their argumentation skills (Guo et al., 2020; Zhang et al., 2017). (See Padilla & Benitez, 2014, for a detailed discussion of theory, relationship to other evidence types, and methods in using processes for validation.)

Once an assessment is deployed, response process data can be a valuable tool for test security and data integrity (Qian et al., 2016). For example, item response times can be used to identify unusual patterns that may suggest cheating (Marianti et al., 2014). Test takers who use a hidden answer key or memorize answers from an illicit examination copy are likely to have a mismatch between their item completion times and the distribution of times from the test-taker population. Such a mismatch may occur in part because, by virtue of not having to engage item solution processes, the former group moves through items more quickly than do honest test takers. Similarly, a test taker who is copying from another test taker will show not only great similarity to that individual's responses, but also synchronicity in timing with those responses. The detection of any such events can allow the possibility of real-time alerts to examination proctors.

Process data can also be used as part of quality assurance to identify when something has gone wrong in administration or scoring. With the addition of technology, quality assurance must include digital delivery, user-interface functionality, data recording, automated scoring, and data transfer and storage. Increasingly, assessment developers look to the best practices of software engineering for proper quality assurance methodologies because the assessment is a software product. Similar to how software engineers create unit tests during development, assessment developers can specify constraints on the expected response patterns while they craft the items. Cognitive labs can also be used to generate estimates of those patterns. During both pilot testing and final deployment, large deviations from the expected patterns would then be flagged for investigation.

After the completion of a test administration, process data can be used to understand performance in greater detail than is possible through test scores alone. Such data can, for example, illuminate group differences in scores (e.g., Bennett et al., 2021; Guo et al., 2019; Zhang, Bennett et al., 2019). Greiff et al. (2015) used process data from PISA TBA items to uncover inquiry strategy use and differential strategy use by country. Such understanding might lead to adjustments to scoring practices (when it is found that scoring privileges one type of solution), teaching (when important solution approaches are not being taught), or instruction (when it is found that a population group is not benefiting sufficiently from existing teaching practices).

At the individual level, process data could be used for formative feedback to the teacher or test taker. Those data might be used to identify test takers who are likely to be guessing (Lee & Jia, 2014), as well as those who appear to hold particular misconceptions or procedural weaknesses. In a more open-ended problem space, the level of depth at which an individual is engaging might also be described. Such evidence might enable identification of not only those who need additional guidance, but also those who could benefit from further challenge.

Analysis and Scoring of Response Process Data

The simplest process data consist of response times augmenting the raw item response. Thus, for every item, the response, its score, and the total time taken are available for analysis. Models of test speededness (van der Linden, 2017) that account for student changes in strategy due to time constraints and models that predict guessing and cheating behaviors have been applied to such data (Guo et al., 2016; van der Linden & Lewis, 2015). The remainder of this section will deal with more complex-response interactions, leading to more data available for each item response.

The rich and varied information contained within complex-response process data presents a significant challenge for analysis. The log file contains an abundance of low-level events that do not readily translate into relevant inferences. Consequently, the statistical methodology used to make those inferences needs to be well understood and appropriately validated. Here, we discuss a few of the more common approaches, along with recommendations for analysis and modeling of process data from complex tasks.

Because of the scale and complexity of the data, psychometric methods traditionally used for the analysis of item responses and test scores are frequently inappropriate. Process data are irregular in that the recorded events not only vary in number and meaning across students but also are context dependent. Furthermore, the parameter space of models used in the analysis of such data is large and assumptions of conditional statistical independence are clearly violated. For these reasons, new methods from machine learning or computational statistics may be better suited to these data. Rather than relying on models that predefine the relationships among variables, in machine learning these relationships are derived from the data. This derivation, or "learning," requires a large amount of data; depending on the complexity of the algorithm and the number of parameters, data requirements range from thousands to tens of thousands of records. These methods are extremely useful when there is a large quantity of factors (variables or parameters) or when the relationships between the factors are complex and ill-defined. In either case, predefining a full model is impractical.

Machine learning methods are either *supervised*, which means that a labeled data set is used to train the models, or *unsupervised*, in which patterns are identified within the data without prior labeling. Labeling presents a particular difficulty for educational applications. While unambiguous classification may be more typical of some traditional machine learning domains, such as computer vision (either the picture contains a cat or it does not), "ground truth" is less common in educational and psychological

assessment. Human raters are frequently used to label data, but allowances should be made for their known fallibility (e.g., see Ho & Kane, 2013, with respect to the rating of teaching processes). Performance replays can be constructed from the process data to aid in human labeling (R. Baker & de Carvalho, 2008), which may improve consistency. Unsupervised learning is used primarily in exploratory analyses as a method for identifying common patterns within a large data set, frequently involving clustering or dimensionality reduction. For assessment purposes, such analyses can be particularly helpful in discovering response processes that are different from the expected approaches.

Both types of machine learning expect input data in which each record is represented as a feature vector—that is, a list of variables having numeric values (see also Shermis et al., this volume). The mapping of raw data to feature vectors, known as feature engineering, is a critical step because anything not encoded into the feature vector will not be usable for classification or clustering. For example, the content of text documents is frequently modeled as a "bag of words" in which the feature vector is simply the count of each dictionary word used in the document. This representation does not include word order, encoding "house boat" identically to "boat house." The representation vector could be expanded by adding bigrams (two-word sequences) or part-of-speech tags to enable such distinctions. For analyzing response process data, features can include the count of specific actions taken, the mean time between actions, or the most frequent action in a given time slice. Feature detectors can be crafted that identify significant patterns of action within the raw data, which can then be added to the feature vector. Once the feature vectors are constructed, machine learning uses a variety of statistical techniques to classify or cluster records.

For assessment, classification can be used to evaluate performance, with categories like "high," "moderate," and "low." Classification can also be used to identify records that are likely the result of guessing or cheating behaviors. For generating formative information from a summative assessment, classification can be used to identify strategies or misconceptions. Machine learning methods applied to the classification of process data include support vector machines, *K*-means clustering, logistic regression, classification and regression trees, and deep neural networks (Baradwaj & Pal, 2011; Rivas et al., 2019).

Common to all machine learning is that few assumptions are imposed about the relationship between the features in the data and the final classifications. The statistical methods iterate to optimize a loss function (e.g., classification match with prior labels, or data fit), but may add and delete factors and relationships between factors or combine the results from multiple models in a weighted fashion. This methodology makes final classifications hard to defend because the logic behind the classification is not transparent.

An alternative, or complementary, approach is theory-driven modeling. In these methods, the relationships between the data and the inference are defined in advance. The models may contain parameters that will be tuned given the data, but the meaning

of the classifications or inferences are clear ab initio. Bayesian networks are an example of modeling complex data with a theory-driven approach. Relationships between the features (observable variables) and the targets of inference (unobservable variables, or latent nodes) are defined by the network structure, while the exact probabilistic relationships between nodes are learned from the data.

More commonly in educational assessment, a combination of the above two approaches is used. Data mining can identify clusters of common action patterns. Content experts then examine these patterns and infer, for example, the strategy that the test takers employed. Once a set of interesting strategies is identified, models can be built to classify records by strategy or score the performance based on the action patterns observed.

One of the more active areas of process data research has been in writing, particularly for essays composed as part of standardized assessment. Using a combination of theory-driven and bottom-up approaches, this research has found meaningful relationships between essay scores and such basic features as the types of pauses that characterize composition and the length of writing bursts (Almond et al., 2012; Guo et al., 2018; Zhang & Deane, 2015). Theoretically predictable differences in feature patterns among writing task types have also been detected (Deane et al., 2018). Studies have used unsupervised data reduction methods like exploratory factor analysis to select and aggregate low-level log file features into scales (Deane, 2014; Zhang & Deane, 2015) and profiles (Bennett et al., 2022). Meaningful differences among writing proficiency levels and among gender, socioeconomic status, and racial/ethnic groups have been discovered using such scales (Bennett et al., 2020, 2021; Guo et al., 2019; Zhang, Bennett, et al., 2019).

For an extended interactive performance, a different approach from modeling a set of extracted features is to model the behaviors of the individuals within the context of the problem that they are solving. Decision models calculate the probability of a person making a choice in a particular situation, given the person's goals and beliefs (C. L. Baker et al., 2011). One can think of this approach as if we were programming an autonomous agent to perform the task. Given a goal, the agent will need to select actions, monitor the results of those actions, and select next steps until the goal is met or it gives up. Partially observable Markov decision processes are one example that can be applied to such assessment performances (Bellman, 1957; Howard, 1960). This model calculates the probability of a person taking a given action in a particular state of the problem as a soft-max[1] over the expected total rewards for taking that action. Goals are encoded into the reward structure, which quantifies both the rewards for reaching different problem states (e.g., a solution to the problem) and the costs of taking specific actions. Beliefs are encoded into the model's transition functions and state space as subjective understanding of both the probabilistic effects of taking particular actions and what is possible. Inferences about the test taker's abilities, goals, and beliefs can be made by fitting the model to the response data produced by the test taker (LaMar, 2018).

Automatically Scoring Complex Constructed-Response Tasks

The preceding section centered on evaluating response processes, a method that could be applied to a wide range of test item types and used for a variety of purposes that do not necessarily factor into the scores reported on consequential tests. In the current section, we focus on scoring per se and on a select class of assessment tasks. In particular, the section addresses automatically scoring a variety of complex constructed-response tasks requiring judgment of product features, process features, or both. We describe the tasks to which such artificial intelligence (AI) approaches have been commonly applied, give a high-level description of how scoring works, and suggest the types of evidence that ought to be considered for validation (see also Shermis et al., this volume).

Bennett and Zhang (2016, p. 142) offered the following definition for automated scoring: "the machine grading of constructed responses that are generally not amenable to exact-matching approaches because the specific form(s) and/or content of the correct answer(s) are not known in advance." As they noted, that definition is quite broad, encompassing grading approaches that differ considerably as a function of the constructed-response task being posed and the character of the answers expected from a given population of test takers.

As of this writing, automated scoring is used operationally by many testing programs, including for postsecondary admissions (GRE General Test Analytical Writing Assessment, TOEFL iBT, Pearson Test of English), occupational and professional licensure (USMLE), and school accountability (selected Smarter Balanced states). The primary motivations are to reduce the cost associated with human scoring and increase the speed of reporting.

The types of tasks to which automated scoring has been applied operationally include essay writing, speaking, architectural design, patient management, accounting, mathematical problem-solving, and relatively short text responses associated with reading a passage or justifying a mathematical problem solution (Williamson et al., 2012). One important dimension along which such tasks may vary is in being static versus dynamic. For instance, in the GRE General Test Analytical Writing Assessment, the product or outcome—that is, the submitted essay response—is the only aspect graded. In contrast, the USMLE includes a section containing 13 computer-based case simulations (USMLE, 2018). As mentioned, each simulation presents a patient management problem that changes as the test taker interacts with it (e.g., the test taker's decision to run a diagnostic test produces a result that must be considered and acted on). Consequently, USMLE automated grading must account for the *process* used to manage the patient, as well as such outcomes as the final diagnosis and prescribed treatment.[2]

Irrespective of the task and AI approach, automated scoring generally includes three conceptually separable parts: feature extraction, feature evaluation, and evidence accumulation (Drasgow et al., 2006). In *feature extraction*, the scorable components of the response are computed (e.g., parsing and tagging words for essay scoring; identifying what actions were—and were not—taken in managing a patient). *Feature evaluation* entails judging the extracted components (e.g., the agreement of subjects and verbs;

the appropriateness and order of actions in patient management). Finally, *evidence accumulation* involves aggregating the feature evaluations to produce one or more scores. For essay evaluation, judgments about subject–verb agreement would be combined with those related to other aspects of essay quality (e.g., organization, development, content); for patient management, evaluations of the appropriateness and order of actions would be combined with those pertaining to the suitability of the diagnosis and treatment.

Within a task type, automated scoring systems can be differentiated along multiple dimensions. One important dimension is whether scores are created to predict human scores empirically or, alternatively, are generated from theoretical propositions or rules. Most approaches to automated essay scoring take the first path. That is, they employ feature weights empirically derived to predict human scores, for example, by linearly regressing those scores on the extracted features (e.g., Burstein et al., 2013). This focus on predicting human scores has been driven by a long tradition of human essay rating, whereby human scores have come to be accepted by many users as a "gold standard" (Powers et al., 2015). An alternative approach used in some scoring systems is to employ propositions or rules based on some theoretical decomposition of what constitutes a quality performance in that task domain. The decomposition is based on expert judgment and may involve having one committee develop the rules and another committee verify the rules and the scores those rules produce, without ever optimizing the automated algorithm to predict human scores per se. That approach was essentially followed in the automated scoring of architectural design problems (Bejar, 1991; Braun et al., 2006) and in medical patient management (Clauser et al., 2016). In such approaches, humans are used as experts in defining scoring features and aggregation rules, as well as for quality control when the scores are produced.

A second dimension along which scoring approaches for a given task type may differ is specific to those approaches that seek to maximize agreement with some criterion, like human scores. The dimension is the extent to which transparent versus black-box machine learning methods are utilized. For example, in some approaches, computable features are developed from a construct theory or from an existing scoring rubric. Once computed, those features are combined by weighting them empirically to produce a score (or other judgment such as a diagnostic categorization). Divulging the computable features and their relative weights permits users to make an evaluation of the extent to which those features cover the construct theory or rubric, aspects in which the features may fall short, and how the weighting comports with the intended construct. In contrast, other approaches to automated scoring extract large numbers of computable features without any prior construct or rubric mapping and algorithmically use whatever features best predict the chosen criterion (usually human scores).[3] Recent uses of large language models (LLMs) for essay scoring would appear to work in this way, though there may be ways to reduce the black-box problem via combination with more interpretable measures (e.g., Mizumoto & Eguchi, 2023). In general, however,

such approaches will not be transparent to users and may not even be scrutable to the system's developers.

The most common approach to evaluating the quality of automated scoring has been to compare the generated scores to those produced by operational human raters under the supposition that machine–human agreement equal to or greater than human–human agreement constitutes validation. As multiple commentators have pointed out, machine–human agreement is best viewed as *one* piece of evidence in the validity argument for automated scores or, said another way, validation should be broadly based (Attali, 2007; Bennett & Zhang, 2016; ETS, 2021; Williamson et al., 2012).

A considerably more comprehensive validation conceptualization has been offered by Bennett and Zhang (2016), which considers multiple sources of evidence. Their conceptualization began with the validity argument for *human* scores. It is necessary to evaluate that argument if human scores are to be used as evidence for validating the automated scores (Bejar, 2012). With respect to human scores, they asked:

- Do the test-taker response processes align with the construct definition? For example, if the computer interface is an unfamiliar one, test-taker time and cognitive resources may be split between figuring out the interface and completing the task, thereby contaminating human scores with irrelevant variance.
- Does the human scoring rubric fully capture the construct definition? If the rubric unintentionally drives human raters toward a subset of that definition, then that subset will dominate scores.
- Are operational human raters using construct-relevant scoring processes? If raters are using shortcuts (e.g., avoiding the extremes of the score scale, using correlates such as response length), then their scores will be a less meaningful evaluation criterion.
- Do raters agree reasonably highly with one another? If not, their justification as a validation criterion will be undermined.
- Do raters treat unusual responses in appropriate ways (e.g., responses that may not fit the existing rubric but clearly indicate a high level of competency; ones that attempt to game their way into a higher score)?
- Do human ratings of one task predict ratings on other tasks from the same universe reasonably well? If not, the argument for an underlying construct will be called into question.
- Do the ratings relate in theoretically predictable ways to other measures of the same construct and to measures of different constructs?
- Do the above results hold to reasonably similar degrees across important population groups? If not, those differences may reflect unfairness in human scoring.

With respect to the validity argument for automated scores, Bennett and Zhang (2016) asked:

- Was the model trained and calibrated on an appropriate sample of artifacts from the target population? If not, it may encounter responses that it cannot properly evaluate.
- Are the model's features related to one another empirically in theoretically meaningful ways and do the features and their weighting fully capture the rubric and construct definition? This question centers on whether the resulting scores are a faithful measure of the intended construct rather than some subset of it or simply a measure of one of that construct's correlates.
- Do the automated scores agree with human ratings (in the best case, with the mean rating taken across multiple experts grading under ideal conditions who agree highly among themselves)? Machine agreement with a consensus among experts makes for a more reliable and arguably more valid criterion than agreement with a single rating generated under more rushed, operational conditions.
- How effectively does the automated scoring handle unusual responses?
- How well do the automated scores predict performance on other tasks from the universe?
- Do the scores relate to external criteria in the expected ways?
- Are the functional characteristics described above invariant across population groups?[4]
- What are the likely intended and unintended impacts of using automated scoring on the behavior of test takers and those who educate them? For example, do students indiscriminately use more low-frequency words and complex sentences (when simpler vocabulary and constructions might better serve given writing purposes) because they believe such use will increase their automated scores?
- How does automated scoring compare on each of these dimensions to human scoring? Notable differences in functioning between the methods should be investigated and explained because they are likely to point to sources of inaccuracy, irrelevant variance, or unfairness in one or the other method.

In closing this section, we offer several important points. The first point is that it is generally not the automated scoring engine being evaluated, but the scores it produces. Those scores depend on the nature of the examination questions posed and the population assessed. The validity of scores may vary to the extent that either questions or population characteristics change.

A second point is that, as noted, agreement with human ratings is questionable as the sole criterion for automated score validation. This statement especially holds if the validity argument for the human ratings themselves has not been firmly established. Rather, multiple sources of evidence should be sought to permit a more rigorous and complete evaluation of the validity argument for automated scores. To the extent that those evidentiary sources suggest the same positive conclusion, the argument will be strengthened.

Third, one of those sources of evidence should be an analysis of the features used in scoring, their weighting, and the alignment of these features and weights to the construct definition. In other words, it is difficult, if not impossible, to rigorously evaluate the validity of automated scores without a clear understanding of how those scores are generated and how that generation method comports with the intended construct. Such an understanding presumes transparency of the automated scoring system. The need for such transparency is being increasingly recognized in legal frameworks governing AI systems used more generally to make decisions that have significant impact on people's lives (G20, 2019; Madiega, 2019; OECD, 2019). That recognition is, in turn, fueling efforts to create explainable AI (Dwivedi et al., 2023; Futia & Vetro, 2019; Kuang, 2017; Maglieri & Comande, 2017; Turek, n. d.).

A final point is that our efforts to build a strong validity argument for constructed-response scoring, whether automated or human, should ramp up as the consequences associated with test results increase. This stipulation holds even when those constructed-response scores are not the major portion of the test. Such is the case because, when the decisions emanating from test results have significant and hard-to-reverse impact, it is the testing program's responsibility to ensure that, to the maximum extent practicable, all test components are meaningful and fair indicators of proficiency.

ASSESSMENTS USED TO SUPPORT INSTRUCTIONAL DECISION-MAKING

The section "Assessments Used to Support Consequential Purposes" dealt with consequential TBA, where decisions based on a single result may have a dramatic influence on an individual, group, or institution and may not be easily reversed. In the current section, we focus on TBA uses that generate more easily reversible decisions with less dramatic effect in any given instance. In particular, we discuss the current landscape for TBAs used to support instructional decision-making in real time at the individual student level. We will briefly trace some of the important milestones in the development of such embedded, technology-based formative assessment, noting relevant innovations in design, task type, scoring, and modeling.[5] (See also Brookhart & DePascale, this volume, for a discussion of formative assessment.)

The use of technology-based assessment for instructional decision-making is premised on the principle that individualizing instruction, as would be done in one-to-one tutoring, is more effective than targeting instruction to the group. This principle was implemented through mechanical devices known as teaching machines, which used simple forms of assessment to direct instruction. One of the earliest such machines was created by Pressey (1926, 1927), whose initial purpose was to build a testing device for presenting and scoring responses to multiple-choice questions. Realizing its value for instruction, he incorporated such rudimentary mastery criteria as not eliminating a question until it had been answered twice correctly. Following publication of Skinner's (1958) seminal "Teaching Machines" article, the same basic ideas led to CAI, which

afforded more complex branching possibilities based on the student's prior response(s) (Suppes, 1972; Van Meer, 2003).

The 1980s saw the advent of intelligent tutoring systems (Sleeman & Brown, 1982), today more commonly called adaptive learning or personalized learning systems. Whereas teaching machines, and the CAI implementations that succeeded them, were built on behaviorist conceptions of learning, intelligent tutors brought cognitive learning models to bear and combined them with AI approaches (Anderson et al., 1985). Those domain-specific cognitive models and AI methods were used to track a student's knowledge state dynamically as it evolved during an instructional sequence and progressively adjust instruction as a function of that changing knowledge state. This real-time assessment was initially deterministic, meaning that uncertainty was not factored into the evaluation of the student's state. These deterministic systems were followed by ones that incorporated probabilistic modeling, typically in the form of Bayesian networks, into their real-time assessment of student knowledge state (Corbett & Anderson, 1995; Mislevy & Gitomer, 1995; VanLehn & Martin, 1998).

Many of the intelligent tutors created over the past several decades did not go beyond the prototype development stage and, consequently, were not widely used in classroom settings (Shute & Zapata-Rivera, 2010). One tutor that has been widely used is Carnegie Learning's MATHia, a direct descendant of Anderson and colleagues' extensive, long-term research program at Carnegie Mellon University (Anderson et al., 1985, 1995; Pane et al., 2013; Ritter et al., 2007). A second widely used tutor is ALEKS (Assessment and Learning in Knowledge Spaces), based on the research of Falmagne and associates at the University of California, Irvine (Doignon & Falmagne, 1999; Falmagne et al., 1990). Both ALEKS, now a product of McGraw–Hill Education, and MATHia are used at the school level as well as in higher education. A third, more recently developed example is Woot Math, which helps students in Grades 3–8 learn core math concepts, starting with rational numbers (Milne et al., n.d.). Rather than the more elaborate cognitive-domain modeling approach taken in MATHia, Woot Math concentrates selectively on a small number of ideas and misconceptions identified as key by expert teachers and researchers. Of note is that, in addition to selecting tasks based in part on Rasch models, the system adaptively determines how fast each student should move along the instructional sequence, offers help based on Bayesian models of each student's understanding, presents additional instructional modules, and inserts new levels with review tasks.

A final example can be drawn from the class of tutors known as personalized learning apps, which can be accessed on mobile phones, tablets, or conventional computers. As an instance from this class, the Duolingo language learning app (https://www.duolingo.com/) offered as of this writing instruction in three dozen or so languages, with the more common language courses having on the order of 150-200 brief lessons. Lessons are organized around topics common to language learning (e.g., shopping, food, entertainment), into which new vocabulary and linguistic structures are progressively integrated. Each lesson is composed of an optional synopsis of the content to

be covered and exercises taking various forms (e.g., matching, selecting and arranging words to form a sentence, speaking a presented sentence). Adaptivity and such learning sciences' principles as spaced repetition are employed in exercise selection via probabilistic models (Protalinski, 2020; Settles & Meeder, 2016), although the particular models used operationally have not been identified.

Intelligent tutors, and their CAI and teaching machine predecessors, had in common that instruction was generally built around the presentation and solution of a series of discrete problems designed to evaluate student standing. In contrast, educational games and simulations often present a more thematically integrated problem-solving experience as a means of facilitating and monitoring learning. That experience may require more problem-solving steps, and the state of the game or simulation may change in response to student actions. Thus, educational software based on games, as well as on simulations, is often more performance oriented.

Like intelligent tutors, games and instructional simulations can, in principle, be built on cognitive-domain models, use AI, and incorporate probabilistic methods to estimate skill level. Estimation can be dynamic, allowing real-time modulation of the state of the problem situation in play or of the difficulty level of the problem situation to be presented next. As of this writing, we could locate no commercially available, widely used, and well-documented examples that possessed all these features. Several commercially available games and simulations, however, do possess one or more of these attributes.

As an example, Math Garden allows students to practice basic addition, subtraction, multiplication, and division skills (Klinkenberg et al., 2011). The game is in many ways quite traditional, presenting a series of drill-type math problems for which correct solutions add garden flowers and garner coins for buying virtual prizes. In addition, no cognitive domain model or AI is employed. Presentation is adapted at the item level using estimates of student competency that incorporate both speed and accuracy. Estimates are dynamically generated via the Elo (1978) rating system (originally developed for chess competitions).

An example from science is Inq-ITS (Gobert et al., 2018; https://www.inqits.com/). This system employs a collection of laboratory simulations that allows students to design and conduct experiments. Inq-ITS uses the Next Generation Science Standards as its cognitive-domain framework, coaching students through the NGSS practices of designing investigations, using evidence to make claims, and backing up those claims with evidence and reasoning. The system employs machine learning techniques to evaluate student responses, generate real-time alerts for teachers, and give students feedback on specific aspects of their inquiry practice. Bayesian knowledge tracing, described in the next section, is employed to estimate student skill levels (LaMar et al., 2017, pp. 143–145).

The formative assessment designs, tasks, scoring, and modeling methods used in intelligent tutors, games, and simulations can run from simple to complex. Common to most systems of interest are presenting a sequence of problems created to facilitate learning some set of domain competencies, scoring in real time, probabilistically

estimating knowledge state (or skill level) from that scored performance, and using that estimate to in some way modulate the current task or choice of the subsequent one. In the following paragraphs, we briefly discuss similarities and differences among these systems in the problems employed, presentation design, response requirements and evaluation, and feedback.

As noted, the problems that these systems present may be discrete or part of a thematically related sequence. Thematically related sequences are often introduced by a scenario intended to give a real-world context, motivate the student, and indicate a goal to be achieved. One or more problems typically follow. In an intelligent tutor, a problem statement, related stimuli, and any associated tools are available to the student. A game would add mechanics (i.e., rules and procedures for play) and such motivating elements as points for correct responding, graphics, and audio. In educational simulations, the student interacts with an invented milieu that is intended to represent the key features of some real-world environment, such that the invented environment mimics the real one. The simulated environment may include runnable models that process inputs and produce outputs like their real-world counterparts (e.g., behaving like a patient with a specific medical condition). Problems may also be presented outside the simulated environment, in which case the student responds to those problems based on his or her interactions with the simulation. Alternatively, or in addition, the students' interactions with the simulation may constitute evidence for evaluating knowledge state. Regardless, the simulated environment itself becomes part of the problem.

With respect to presentation, in intelligent tutors, presentation is typically adaptive. Games and simulations, however, may have linear, item-level adaptive, or multistage adaptive designs. In linear designs, all students encounter the same problems in the same order. Many educational games take this approach, allowing a student to progress to a higher level only when some criterion performance has been reached (e.g., a fixed number of problems answered correctly). In item-level adaptive designs, the difficulty of the next problem is determined at least in part by performance on earlier problems. Multistage designs differ from item-level adaptive ones primarily in the frequency of adaptation, adjusting the difficulty of tasks at the stage level, rather than at the item level.

As to response requirements and evaluation, in intelligent tutors, games, and simulations, such requirements may involve entering a number, clicking on a hot spot, moving objects on screen, writing text, or manipulating sliders, dials, or other components. Responses in all three types of systems can be evaluated in terms of the resulting product, the process used to generate that product, or both, depending on the underlying domain model. However, even when the domain model specifies that evidence of proficiency lies in the correctness of an outcome (e.g., a mathematical result), examining the process used to generate that outcome can be the basis for action. That action could take the form of feedback (e.g., providing hints, pointing out errors in process, showing a worked example), choosing new problems, or making other adjustments to instruction.

Feedback itself may be provided before or after responding. A response that is not given quickly enough may trigger some event to occur. That event could be visible, as in the provision of a hint, or invisible, as in updating the estimation of the student's knowledge state. Feedback may also immediately and visibly follow an action—for example, an English-language-learning game character might offer a puzzled look if a grammatically incorrect sentence is entered.

Probabilistic Models for Assessment Embedded in Instructional and Learning Systems

As indicated in the section immediately above, to provide feedback, as well as to adjust instruction, the formative assessment embedded in the learning and instructional systems of interest uses probabilistic models. In this section, we explore some of these models, their evaluation, and their validation.

Within tutoring systems, "student models" are used to track relevant parameters for various characteristics of interest including proficiency, engagement, and affective state (Johns & Woolf, 2006). These models are essential for enabling personalization of the tutoring experience, with the most basic model indexing where the student might be in relation to mastering the content being taught. In such models, proficiency is estimated dynamically from the student's performance to enable real-time adjustment to instruction. Estimation is done through some form of psychometric (or other probabilistic) model.

Although adaptive tests also dynamically estimate proficiency, they presume no change in proficiency over the course of the assessment. In contrast, proficiency should be expected to change because of interacting with a tutoring system. In addition, the system must be able to generate frequent formative feedback. Effective feedback requires specific, accurate, actionable observations about student performance (Hattie & Timperly, 2007; Shute, 2008). Thus, models need to be able to estimate specific strengths, weaknesses, and, in some domains, misconceptions.

For adaptive learning systems like intelligent tutors, the purpose of the student model extends beyond direct feedback to enabling the selection of appropriate next-instructional steps. These interventions might be in the form of hints, encouragement, explanatory text or video, a worked example, or selection of the next problem.

At a high level, the student model can be decomposed into the representation of the student's state and the method used for updating that representation. Approaches to representing student state vary considerably, but some common ones include overlay models, which characterize the student's knowledge as a subset of an expert's knowledge; perturbation (or buggy) models, which catalog both the student's correct and incorrect ideas or misconceptions; and stereotype models, which represent the student through membership in predefined classes (Chrysafiadi & Virvou, 2013; Desmarais & Baker, 2012). The methods for estimating and updating these models also vary widely, but frequently they involve a combination of a hidden Markov model (HMM) and a more traditional psychometric model.

The dynamic nature of updating estimates of student state makes the HMM a particularly suitable statistical tool. The HMM (see Figure 9.5) specifies how an unobserved latent variable, for example, mastery of skill *K*, might change over time based on observed variables, x_i, for example, answers given to problems within the learning system. The statistical relationship between the unobserved variable and the observed one is known as the emission probability, $p(x_i = 1|K_i)$. For many student models, this emission probability takes the form of a psychometric model (e.g., IRT, cognitive diagnostic model, CDM; Rupp et al., 2010). The probability of the latent variable changing between student actions is known as the transition probability, $p(K_i = 1|K_{i-1})$. Within adaptive learning systems, this probability is related to the learning rate. Additional parameters, frequently represented as ξ, can include amount of time between occurrences and a quantification of the intervening instruction.

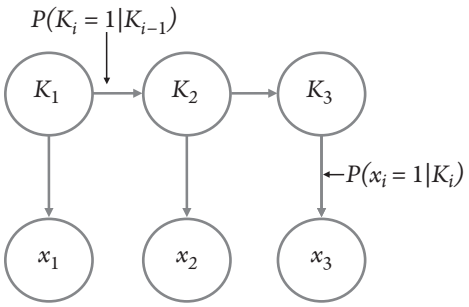

FIGURE 9.5
A Hidden Markov Model for Estimating Student State Over Time

Because the form and content of adaptive learning systems vary widely, the modeling approaches also vary. For learning systems that primarily present a series of problems for the student to work through (e.g., ASSISTments, Heffernan & Heffernan, 2014; Andes, vanLehn et al., 2005), the key inference at any given moment is whether the student has mastered a particular, narrowly defined skill. This inference allows the tutor to introduce a new skill once the current one has been mastered. Thus, an overlay model is frequently employed to represent the student, with a discrete set of Boolean variables tracking mastery of the skills or knowledge components of interest. The most common updating method for this purpose is Bayesian knowledge tracing (BKT; Corbett & Anderson, 1995), which takes the form of an HMM. In BKT, the probability of a student having mastered a skill *L* on their *n*th attempt is modeled as the sum of the probability they had previously mastered it at the $(n-1)$th attempt and the probability that they had not previously mastered it but now have mastered it (transitioned) on this attempt (T_n):

$$p(L_n) = p(L_{n-1}|X_{n-1}) + (1 - p(L_{n-1}|X_{n-1})) * p(T_n) \quad (1)$$

The emission probabilities in BKT usually take the form of a CDM (Rupp et al., 2010), which models correctness of response based on a set of skills represented as Booleans, either mastered or nonmastered. CDM models include slipping and guessing parameters to account for an incorrect response when the requisite skills are present (slipping) and a correct response when the requisite skills have not been mastered (guessing). Combinations of skills can be modeled either as conjunctive or as compensatory, allowing the flexibility to apply appropriate cognitive theory for the domain.

Another common approach to student modeling within adaptive learning systems is performance factor analysis (Pavlik et al., 2009). Such models characterize student knowledge in terms of continuous variables and make use of an updating mechanism that is a variation of IRT with a learning parameter (slope).

More recently, advances in recurrent neural networks have been applied to the problem of estimating and updating student proficiency variables in what is known as deep knowledge tracing (DKT; Piech et al., 2015). DKT uses the same inputs and outputs as BKT, but rather than computing probabilities of skill mastery and using them to estimate subsequent performance, the probabilities of success on future items are predicted directly with a recurrent neural network, usually using a long short-term memory layer to ensure that past performance has an extended impact on future predictions. At the time of this writing, several extensions to DKT have been developed that augment the input vectors with both item information (including difficulty and skill tapped) and additional performance information, such as the amount of time taken to solve the item (Ai et al., 2019). Structural improvements to the predictive models have also been proposed, including applying regularization to prevent large variance in the predictive output (Yeung & Yeung, 2018). While the recurrent neural networks approach requires a large training data set, the flexibility of these models to utilize complex inputs, incorporate domain-specific data transformations, and interface with other models makes it likely that DKT and other uses of deep neural networks will have a major impact on adaptive learning technology in the near future.

Models that are more deeply grounded in learning theory have also proven useful for interpreting behavior in solving complex tasks. In particular, the ACT-R cognitive architecture (Anderson et al., 1997) has been used to predict student actions in multistep problem-solving using a technique known as *model tracing* (Corbett et al., 1995). ACT-R models cognition at a fairly low level, including declarative memory, working memory, and procedural knowledge in the form of if–then production rules. These models predict not only keystroke-level student actions, but also the time between actions, providing additional predictions that can be used for model validation. Because there are so many parameters involved in an ACT-R model, individual parameter estimation is not attempted. Instead, student actions are compared to an ideal student, or expert, model. In some adaptive learning systems, this comparison is sufficient because any off-path action triggers tutoring until the student returns to the ideal path.

Bayesian models and estimation have been popular in adaptive learning systems, including the intelligent tutors built on ACT-R (Koedinger & Corbett, 2006). One

attraction is that Bayesian estimation can occur iteratively as each new data point updates the models' probability distributions. This real-time updating allows the system to act on an estimate as soon as it passes a set certainty threshold. In ACT-R tutors, this "knowledge-tracing" technique allows the system to select tasks that exercise knowledge components the student is not likely to have already mastered (Aleven et al., 2017).

One particular type of Bayesian model, the Bayesian network (Bayes nets), is frequently used to represent complex relationships between the evidence emerging from student performance and the latent variables that make up the student model (Chrysafiadi & Virvou, 2013). A Bayes net is a directed acylic graph in which nodes represent variables and relationships between nodes are described with conditional probability distributions. In adaptive learning systems, the variables represented are almost always categorical, making these relationships conditional probability tables (CPT).

Figure 9.6 shows a simple Bayes net that might be part of a psychometric model (note that there are too few observable variables for this model to be identifiable given the number of latent variables). The observable variables (x_1, x_2, x_3) are indicators (behaviors) from a student performance, while K_1 and K_2 are latent variables that explain differences in performance. In this example, we include an intermediary latent node S_1 that could indicate a particular strategy students might implement. The observable behaviors x_1 and x_2 are indicators of use of this strategy and, as such, are not conditionally independent given K_1, but are conditionally independent given S_1. Observable x_3 has two parents (K_1, K_2) making this variable a within-observable multidimensional measurement model. The way in which K_1 and K_2 interact to produce the observable x_3 can be flexibly defined by the CPT of x_3, allowing for compensatory, noncompensatory, and more complex relationships to be modeled.

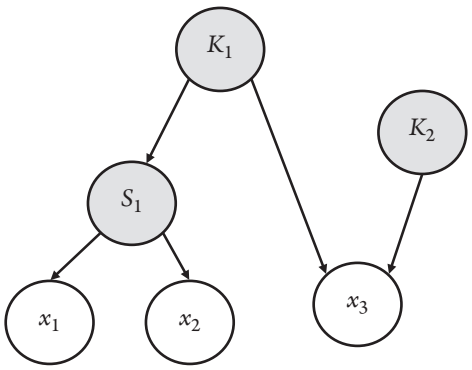

FIGURE 9.6
A Simple Bayesian Network
Note. The gray nodes represent latent variables and the white nodes are observed variables.

A Bayes net can be updated when new observations are made, giving a running dynamic probability distribution for the latent variables. One of the most attractive properties of using Bayes nets for student models is that complex relationships between cognitive and affective latent variables can be modeled to explain specific behaviors. Thus, for learning systems that employ dynamic estimation from student interactions in a simulation, game, or other complex performance, the Bayes net is a natural choice. These models can also be found in formative assessments that include both complex evidence and student models (Almond et al., 2007, 2015).

Evaluation of Student Models in Instructional and Learning Systems

The overall functioning of adaptive learning systems has generally been evaluated based on the achievement gains produced as compared to a traditional curriculum (e.g., Pane et al., 2013). While such gains may suggest that the system is working properly, they do not necessarily confirm the validity of the student models. Similarly, the absence of learning gains does not necessarily imply a failure of the models. Because the function of the student model is to enable feedback or adaptation of tutoring, the model's value might best be judged with respect to how effectively it achieves these goals during the learning session. Experimentally contrasting a linear presentation of problems with one dictated by knowledge tracing, for example, can identify the student model's impact on achievement (Corbett et al., 2000, as cited in Aleven et al., 2017, p. 528).

Other evidence as to the quality of model functioning can be provided by how well the models predict the student's next action or the success of their next solution attempt. Because this prediction is seen as a classification task, researchers in the learning system and data-mining communities have frequently used a classifier accuracy metric known as area under the curve. Area under the curve is calculated by plotting the probability of a true positive by the probability of a false positive over the range of possible threshold values for the classifier. The area under the curve would thus be 1.0 for a perfect classifier and 0.5 for random selection.

COMBINING CONSEQUENTIAL DECISION-MAKING AND INSTRUCTIONAL SUPPORT: CAN TECHNOLOGY-BASED ASSESSMENT DO BOTH?

Earlier sections addressed assessments used to support consequential purposes and assessments used to support instructional decision-making. The affordances of technology make it easier to imagine the possibility of generating information from the same assessment that might simultaneously serve consequential decision-making and instructional-support purposes. Several approaches might be taken to satisfy this dual end, with the approaches differing in fundamental ways, especially in foregrounding one or the other purpose.

The least radical approach would entail generating formative information from consequential TBA. Here, the consequential purpose would be central and the formative purpose supplementary. As an example, a state English language arts accountability assessment will typically produce total scores, proficiency levels, and other scores. One or more of the other scores may come from a response to an essay task, including scores summarizing overall writing quality as well as ones characterizing such writing traits as content and conventions. To be sure, such trait information provides only a very minimal level of formative information. More valuable might be a description of *how* the student wrote the essay, generated by capturing and analyzing the keystrokes used to create it (Zhang, Bennett, et al., 2019). For instance, was there evidence of planning (e.g., the rudiments of an outline early in the writing session), verbal fluency (e.g., long bursts of text), word-level monitoring (e.g., correction of typos), or global monitoring (jumps from one cursor position to another followed by insertions and deletions)? Were there indications of difficulty in typing (signaled by consistently lengthy intervals between key presses) or of problems in word finding (e.g., long pauses between words)? Possibly more valuable for the teacher and student would be the ability to replay, in speeded-up time, the essay as it was written, thereby giving an opportunity for discussion and reflection on the writing process (Vandermeulen et al., 2020). A process description and replay could lead to the realization that typing practice was needed or that instruction in planning or in how to edit should be provided.

The above approach clearly foregrounds consequential assessment by building into it secondary mechanisms for more effectively supporting instructional decisions. Those mechanisms will be far from ideal. Such feedback will occur late in the school year, be limited to that which can be either wrung from the existing assessment or designed into it without compromising its primary purpose or practicality, and be divorced from the context of classroom instruction. Along with other reasons, these limitations have motivated third-generation proposals that attempt to better balance the two assessment purposes. That is, these proposals move toward foregrounding instructional support decisions, with information for consequential decisions coming as a by-product. Periodic, or through-year, assessment offers one version of this idea (Bennett, 2010b; Bennett & Gitomer, 2009; Northwest Evaluation Association, 2019).

In through-year assessment, measures are repeated at various time points. A key benefit of this model is that consequential decisions (e.g., school accountability) would no longer be based on a single measure taken at one time point, but rather on some aggregation of the through-year observations. Also, more frequent feedback would be provided from each of the through-year events.

Many features of this idea could in principle be varied, including the number of TBAs administered, what content they cover, when they are given, the extent to which districts and schools can choose the ordering, and how results are accumulated into a performance index. The greater the variation in these features across schools or districts, the lower will be the comparability of the results (Bennett, 2020). The highest level of comparability would generally result from administering a fixed number of assessments

created from the same content specifications at selected time points to each student under the same conditions and scored with the same aggregation rules. Administering different forms of a comprehensive interim assessment during 1-week windows in fall, winter, and spring would be a simple example, though the extent to which it would foreground instructional support decisions might be arguable (Shepard et al., 2011). More attuned to instructional support decisions would be a regimen in which each assessment focused on a different critical competency associated with grade-level standards. Such a design would either require all schools to follow the same instructional sequence or allow them to choose the order of administration, complicating comparability and test security. (See Georgia Department of Education, 2018, pp. 80–82, for a description of such an assessment piloted under the federal Every Student Succeeds Act Innovative Assessment Demonstration Authority.)

In through-year models, many rules for aggregating results are possible, each derived from a different conception of achievement (Wise, 2011). Weighting the results as a composite with the greatest emphasis given to the most recent assessment would reflect a view of achievement as competency accumulation. A different weighting could mimic the way grades are averaged over quizzes, midterms, and final exams, with this more even weighting suggesting a view of achievement as a collection of accomplishments. A third possibility is to aggregate results based on test information, which would emphasize precision in measuring the dimension common to the several assessments. Finally, taking some function of the difference between the first and last result conceptualizes achievement as the extent of growth. Psychometric models related to these different types of aggregation have been explored by Mislevy and Zwick (2012), Fu and colleagues (Fu & Feng, 2018; Fu et al., 2012, 2013), and Rijmen (2009).

Perhaps the most radical proposal for foregrounding instructional support decisions while generating information for consequential decisions as a by-product is to use a record of daily learning interactions, potentially removing the need for separate assessments entirely (Bennett, 1998, pp. 11–14; Gee & Shaffer, 2010; Pellegrino et al., 2001, pp. 283–287; Tucker, 2012).[6] There are many attractions to this third-generation idea, including the sociocognitive one that the contexts for learning and assessment become identical. That is, both activities draw on the very same content, knowledge representations, and tools (Bennett, 2015), increasing the value of the derived information for adjusting instruction. An example of how this idea might work instructionally can be found among adaptive learning systems like MATHia (Ritter et al., 2007). As noted, these intelligent tutors estimate from real-time learning interactions what a student knows and can do with respect to a content domain and then base moment-to-moment adjustments on that dynamically computed estimate.

Using this type of data for consequential decision-making, however, would be very challenging (Bennett, 2015). One obvious concern is that the assessment model is essentially an extreme case of through-year assessment—that is, one without constraints. The reality is that, within a state and even within many districts, students in the same grade use *many* different electronic learning applications. These applications

vary in their content coverage, emphases, and order; knowledge representations and tools; rigor and quality of problems posed; and scoring criteria and methods, among other things. Further, some environments are likely to be employed more frequently with certain demographic groups, thereby confounding group and measure. A related, but broader concern is that the evaluation instrument is no longer independent of the school or district, leading to the perception that these entities are appraising themselves. Also possibly problematic is the continuous recording of behavior for use in consequential decision-making, which raises both privacy concerns and such educational ones as the potential for discouraging experimentation in teaching and learning. Discouraging experimentation would be unfortunate because such behavior serves a critical function in learning (Kapur, 2010).

Several possibilities exist for reconciling more effectively the goals of consequential assessment with those of assessment for instructional support. These possibilities revolve around systems of assessment—that is, TBAs designed to function synergistically in their pursuit of different goals because trying to fashion a single method for achieving competing goals leads to a suboptimal solution for each goal. One such proposal follows the competitive sports model (Bennett, 2015), a variation of through-year assessment. During instruction, students utilize whatever electronic and other learning environments their schools employ, with learning interactions recorded and used for guiding (and occasionally describing) instruction but never for consequential purposes (as in the practice periods before, and often interspersed within, a sports competition). Students and teachers are informed when an assessment for consequential purposes is to occur (i.e., the actual competition). That assessment is common in design, content specifications, administration window, and scoring across all schools and districts. For state policy makers, such a system could provide the data needed to evaluate how well individual schools were performing in educating all groups of students (from some aggregation of the common assessments' results), along with descriptive data about what students in each school were doing (from sampling the recordings of learning behavior). The latter data might allow instruction to be described at an unprecedented level of detail, greatly enhancing our understanding of learning activity, content, and rigor differences occurring across teachers, classes, schools, districts, and demographic groups.

SUMMARY AND CONCLUSION

This chapter considered three broad classes of TBA: ones used to support consequential purposes, ones used to support real-time instructional decision-making, and third-generation ones that attempt to combine both purposes.

For consequential testing programs, the rationale for moving to TBA is tripartite: align the testing medium with that of learning and of the information economy workplace, conduct assessment processes more efficiently, and measure what previously could not be measured as well or at all. In the United States, a first-generation

infrastructure exists for achieving these goals. Many state assessments, national and international group-score assessments (NAEP, PISA, Programme for the International Assessment of Adult Competencies [PIACC]), graduate and professional admissions measures (GRE General Test, TOEFL iBT), and occupational and professional assessments (USMLE, Architect Registration Examination, Praxis) are routinely delivered by computer in educational institutions or in dedicated centers. Other assessments, most prominently college entrance examinations, are beginning to be delivered in that mode as well.

As a result of the COVID-19 pandemic, some consequential programs have moved to include delivery via remote human and AI proctoring to test takers' homes and offices. Such administration had been used successfully on a small scale in a few niches for several years (e.g., competency-based education at the university level and, more recently, English language assessment via the Duolingo English Test). The COVID-19 pandemic, however, greatly accelerated testing at test-taker locations. In 2020, the GRE General Test, TOEFL iBT, Praxis, HiSet (a high school equivalency examination), the Law School Admission Test, and the International English Language Testing System added home options. Whereas there has been relatively little published research on the technical quality of assessments so delivered, we expect this evolution to continue as need drives use, with research catching up.

The successful deployment of a robust first-generation delivery infrastructure for consequential testing has offered a foundation for innovations in assessment design. Those innovations generally aim to increase construct fidelity by evoking more complex cognitive processes and allowing responses to be observed in finer detail. Four kinds of innovative assessment design were distinguished: TEIs, extended-interaction items, scenario-based tasks, and simulation-based performance tasks. Gathering validity evidence regarding the functioning of these types is critical to asserting that the evoked processes are in fact the intended ones, that the innovations do not introduce unfairness for groups or individuals, and that the Person × Task interaction associated with scenario-based and simulated-based performance tasks is appropriately accounted for.

The changes to practice that TBA entails inevitably raise challenges for score comparability. These challenges occur because many programs offer tests in both paper and technology modes, other programs wish to maintain TBA score continuity with past paper versions, and still other programs want to ensure constancy of score meaning when the digital test is offered on multiple technology platforms. Different degrees of comparability may be appropriate for different use cases and achievable through different methods. Comparability may be studied, and the data needed to make score adjustments gathered, via various designs. However, technology is changing rapidly, forcing assessment programs to change as well. That rapid evolution may make it impractical to study and adjust continuously for the effects of new assessment implementations. Thus, it may be necessary to replace an empirically focused approach with one grounded in design principles directed at maximizing validity and fairness for individuals and groups. Empirically based methods like those traditionally used for linking may still play a role

as one of the investigative methods used to derive assessment design principles or to document the degree to which device-agnostic results were obtained. However, empirical approaches cannot be relied on as the primary mechanism for ensuring and maintaining comparability. Moreover, the desire to maintain comparability over time needs to be balanced against giving testing programs license to incorporate the innovation required to remain relevant in a rapidly changing world.

One of the as-yet underutilized benefits of TBA is certainly its ability to capture evidence of response processes. Such evidence can contribute to the validity argument for an item or task, help in the identification of guessing or cheating, and indicate strategy use or the presence of misconceptions. Response process data may include actions, resulting events, and latencies. Those data may index relatively basic cognitive processes like fluency or higher level strategies, plans, and knowledge. Because of their complexity and scale, process data are often analyzed and modeled using relatively opaque, bottom-up methods like machine learning, as well as ones from computational statistics that allow for the instantiation of theoretical propositions (e.g., Bayesian networks). Combining bottom-up and theory-driven approaches holds promise in that machine learning can be employed to help build the theory to be implemented in, for example, a Bayesian network.

In contrast to response process data, the automated scoring of complex constructed responses is used operationally in many consequential testing programs. This use is driven by the need to cut costs and increase reporting speed. Such scoring usually focuses on an end product, such as an essay or an architectural design, although process data are also included in some instances (e.g., medical patient management). Multiple approaches can be used to score a given task type. Important distinctions among methods relate to whether the automated scores are created to predict human scores empirically or generated from theoretical propositions or rules. A second distinction specific to approaches that seek to maximize agreement with a criterion like human scores is the extent to which transparent versus black-box methods are employed. Whereas validation of scoring has usually focused on agreement with human scores, such agreement is best viewed as one piece of evidence in a more comprehensive validity argument. That more comprehensive argument may, among other things, need to include an evaluation of the validity of using the human scores as a criterion. Additionally, how automated scores are produced, and the alignment of that method with the intended construct, should be a consideration. Because of concerns over the use of AI in society generally, we should expect significant work on explainable, transparent scoring methods.

The chapter's second major section dealt with assessments used to support instructional decision-making. Such assessment is incorporated into intelligent tutoring (or adaptive/personalized learning) systems, educational games, and simulations. Intelligent tutors typically deliver instruction built around discrete problems, which may be similar to the TEIs and extended-interaction items found in consequential tests, whereas games and simulations often present a more thematically integrated problem-solving experience utilizing performance tasks. Tutors, games, and simulations

may employ some combination of cognitive domain model, AI, and probabilistic method to estimate competency dynamically and, thereby, modulate the current problem or choice of the subsequent one. Competency is tracked through a student model that includes a representation of the student's knowledge state (referenced to the larger cognitive domain model) and a method for updating that representation. Probabilistic methods employed in that updating have included BKT and Bayesian networks. MATHia, ALEKS, and the Duolingo language-learning apps are examples of such tutors in widespread use. Examples of games or simulations that appear to personalize learning via such methods include Math Garden and Inq-ITS.

In tutors, games, and simulations, response evaluation can focus on the product (e.g., answer to a math problem), the process used to generate it, or both, depending on the underlying cognitive domain model. However, even when the model specifies that evidence of proficiency lies in the correctness of a product, analyzing the process can suggest appropriate next steps in the form of feedback, selecting new problems, or making other instructional adjustments.

Used in some commercial instructional applications are automated scoring technologies, most commonly for writing or speaking, that share methodology with those employed for consequential assessment. Examples in composition include MyLab Writing and Criterion. These systems evolved separately from intelligent tutors. As a result, the former systems do not generally use cognitive domain models or probabilistic methods to estimate and represent student competency or to personalize instruction, instead only rating and giving feedback on each response in isolation. We should expect to see systems that use automated scoring to support instruction converge with intelligent tutors because the two approaches offer complementary capabilities.

The last section in this chapter described third-generation approaches that combine instructional support with consequential purposes. Approaches differ in the extent to which instructional support decisions or consequential decisions are foregrounded. Included were generating formative information from consequential TBA, through-year assessment, and generating consequential information from ongoing learning interactions. The last approach, while seemingly attractive, raises questions of comparability, privacy, and the potential for negative educational consequences. An approach based on the competitive sports model was proposed as a path that might offer the desired benefits while mitigating the issues raised by more radical models.

What might be productive directions for research? There are many possibilities and here we suggest but a few. One important direction for both consequential testing and instructional support purposes might be to build the foundation for operationalizing the use of process data. For example, research should be directed at identifying whether diagnostic profiles can be identified suggestive of differential instructional action. Placement in a profile could be part of the formative output from a consequential test (e.g., a writing assessment) or as part of personalized instruction. Questions concern

whether meaningful, distinct classes of behavior can be identified (e.g., dysfluent writer vs. fluent writer), whether placement of individuals in such classes is stable across tasks from the same domain, whether the addition of product information adds value, and whether instruction based on such placement results in more learning than a profile-agnostic alternative.

An associated direction concerns how best to report process data, whether in conjunction with a profile or not. If the reporting goal is to encourage reflection as to how one went about problem-solving, then performance replay should be studied. Key questions include identifying the replay formats that best facilitate reflection, learner action, and change in competency.

A third direction relates to the impact of simultaneously incorporating cognitive domain models, AI, and probabilistic methods into games and simulations. These capabilities are not yet commonly found together in commercial products. Experiments could be conducted to estimate the incremental effect of using such additions to increase personalization. A similar direction could be taken with respect to the insertion of cognitive domain models and probabilistic methods in learning systems that use automated scoring.

A fourth direction is associated with the implications of LLMs for validity, modeling, and analysis. The potential uses of LLMs are wide ranging, including item generation, constructed-response scoring, feedback, and reporting, among other possibilities (Bulut et al., 2024; Hao et al., 2024). This direction is notable because of the great interest evident in the field and the potential for improvements in efficiency and quality it seems to portend. Although this direction may pose new issues, because LLMs are a subcategory of AI, many of the challenges LLMs bring are the same ones as already noted in, for example, the section on Automatically Scoring Complex Constructed Response Tasks (e.g., bias, explainability, the need for validation criteria to be broadly based), as well as in other publications (e.g., Bejar, 2012; Bennett & Zhang, 2016).

Finally, calls to merge the purposes of consequential tests and instructional support are likely to grow as learning activity increasingly occurs online. Research should focus on studying the technical quality and instructional utility of approaches that attempt to account in principled ways for the challenges inherent in combining these divergent purposes. The competitive sports model is one example. Research should attempt to determine whether it can produce meaningful consequential results at the same time as it describes and guides instruction.

ACKNOWLEDGMENTS

We wish to acknowledge Steve Ferrara, Ryan Baker, and Denny Way for their helpful reviews on earlier drafts of the manuscript. Steve's reviews were as he was in life: constructive, gentle, modest, optimistic, and wise. This chapter is dedicated to his memory.

REFERENCES

Ai, F., Chen, Y., Guo, Y., Zhao, Y., Wang, Z., Fu, G., & Wang, G. (2019). Concept-aware deep knowledge tracing and exercise recommendation in an online learning system. In M. Desmarais, C. F. Lynch, A. Merceron, & R. Nkambou (Eds.) *The 12th International Conference on Educational Data Mining,* (240–245). Educational Data Mining.

Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2017). Instruction based on adaptive learning technologies. In R. E. Mayer & P. Alexander (Eds.), *Handbook of research on learning and instruction* (2nd ed., pp. 522–560). Routledge.

Almond, R., Deane, P., Quinlan, T., Wagner, M., & Sydorenko, T. (2012). *A preliminary analysis of keystroke log data from a timed writing task* (ETS Research Report No. RR-12–23). ETS.

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J.-D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement, 44*(4), 341–359.

Almond, R. G., Mislevy, R. J., Steinberg, L., Yan, D., & Williamson, D. (2015). *Bayesian networks in educational assessment*. Springer.

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

Anderson, J. R., Boyle, C. F., & Reiser, B. J. (1985). *Intelligent tutoring systems. Science, 228*, 456–462.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences, 4*(2), 167–207.

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-R: A theory of higher level cognition and its relation to visual attention. *Human–Computer Interaction, 12*(4), 439–462.

Association of Test Publishers. (2017). *Innovative item types: A white paper and portfolio*. Author and Institute for Credential Excellence.

Attali, Y. (2007). *Construct validity of e-rater® in scoring TOEFL® essays* (ETS Research Report No. RR-07-21). ETS.

Backes, B. & Cowen, J. (2018, April). *Is the pen mightier than the keyboard? The effect of online testing on student measured achievement* (Working Paper 190). American Institutes for Research.

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief–desire attribution. *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society, 33*, 2469–2474.

Baker, R., & de Carvalho, A. (2008). Labeling student behavior faster and more precisely with text replays. *Educational Data Mining, 2008*, 2469–2474.

Baradwaj, B. K., & Pal, S. (2011). Mining educational data to analyze students' performance. *International Journal of Advanced Computer Science and Applications, 2*(6), 63–69. https://arxiv.org/ftp/arxiv/papers/1201/1201.3417.pdf

Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. *Journal of Applied Psychology, 76*, 522–532.

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31*(3), 2–9.

Bejar, I. I., & Braun, H. I. (1999). *Architectural simulations: From research to implementation* (ETS. Research Memorandum No. RR-99-2). ETS.

Bellman, R. (1957). A Markovian decision process. *Journal of Mathematics and Mechanics, 6*, 679–684.

Bennett, R. E. (1998). *Reinventing assessment: Speculations on the future of large-scale educational testing*. Policy Information Center, ETS. https://www.ets.org/research/policy_research_reports/pic-reinvent

Bennett, R. E. (1999). Computer-based testing for examinees with disabilities: On the road to generalized accommodation. In S. Messick (Eds.), *Assessment in higher education: Issues of access, student development, and public policy* (pp. 181–192). Lawrence Erlbaum Associates.

Bennett, R. E. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *Journal of Technology, Learning, and Assessment, 1*(1), 2–23. http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1667

Bennett, R. E. (2010a). Technology for large-scale assessment. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed., Vol. 8, pp. 48–55). Elsevier.

Bennett, R. E. (2010b). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8*, 70–91.

Bennett, R. E. (2012). On the meaning of constructed response. In W. C. Ward & R. E. Bennett (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1– 27). Routledge.

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education, 39*, 370–407. https://doi.org/10.3102/0091732X14554179

Bennett, R. E. (2020). Interpreting test-score comparisons. In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability issues in large-scale assessment* (pp. 227–235). National Academy of Education. https://doi.org/10.31094/2020/1

Bennett, R. E. (2024). Personalizing assessment: Dream or nightmare? *Educational Measurement: Issues and Practice, 43*(4), 119–125. https://doi.org/10.1111/emip.12652

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment, 6*(9), 3–38.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K–12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–61). Springer.

Bennett, R. E., Goodman, M., Hessinger, J., Kahn, H., Ligget, J., Marshall, G., & Zack, J. (1999). Using multi-media in large-scale computer-based testing programs. *Computers in Human Behavior, 15*, 283–294.

Bennett, R. E., Persky, H., Weiss, A. R., & Jenkins, F. (2007). *Problem solving in technology-rich environments: A report from the NAEP Technology-Based Assessment Project* (NCES 2007-466). National Center for Education Statistics, U.S. Department of Education. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007466

Bennett, R. E., Persky, H., Weiss, A., & Jenkins, F. (2010). Measuring problem solving with technology: A demonstration study for NAEP. *Journal of Technology, Learning, and Assessment, 8*(8), 3–44.

Bennett, R. E., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 142–173). Routledge.

Bennett, R. E., Zhang, M., Deane, P., & van Rijn, P. W. (2020). How do proficient and less proficient students differ in their composition processes? *Educational Assessment, 25*(3), 198–217. https://doi.org/10.1080/10627197.2020.1804351

Bennett, R. E., Zhang, M., & Sinharay, S. (2021). How do educationally at-risk men and woman differ in their essay-writing processes? *Chinese/English Journal of Educational Measurement and Evaluation*|教育测量与评估双语季刊, *2*(1), 1–17. https://www.ce-jeme.org/journal/vol2/iss1/1

Bennett, R. E., Zhang, M., Sinharay, S., Guo, H., & Deane, P. (2022). Are there distinctive profiles in essay-writing processes? *Educational Measurement: Issues and Practice, 41(2)*, 55–69. https://doi.org/10.1111/emip.12469

Berman, A. I., Haertel, E. H., & Pellegrino, J. W. (Eds.). (2020). *Comparability issues in large-scale assessment*. National Academy of Education. https://doi.org/10.31094/2020/1

Braun, H., Bejar, I. I., &Williamson, D. M. (2006). Rule-based methods for automated scoring: Application in a licensing context. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 83–122). Lawrence Erlbaum Associates.

Breithaupt, K. J., Mills, C. R., & Mellican, G. J. (2006). Facing the opportunities of the future. In D. Bartram & R. K. Hambleton (Eds), *Computer based testing and the Internet* (pp. 219–251). Wiley.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25*(1), 27–40.

Brown, T., Chen, J., Ali, U., Costanzo, K., Chung, S., & Ling, G. (2015). *Mode comparability study based on PARCC Spring 2014 field test data* [Unpublished manuscript]. ETS.

Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., Ormerod, C., Fabiyi, D. G., Ivan, R., Walsh, C., Rios, O., Wilson, J., Yildirim-Erbasli, S. N., Wongvorachan, T., Liu, J. X., Tan, B., & Morilova, P. (2024). The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. *Chinese/English Journal of Educational Measurement and Evaluation* | 教育测量与评估双语期刊, *5(3)*. DOI: https://doi.org/10.59863/MIQL7785

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 55–67). Routledge.

Chrysafiadi, K., & Virvou, M. (2013). Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications, 40*(11), 4715–4729.

Clauser, B. E., Margolis, M. J., & Clauser, J. C. (2016). Issues in simulation-based assessment. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 49–76). Routledge.

College Board. (2024, January 4). *Everything you need to know about the digital SAT*. https://blog.collegeboard.org/everything-you-need-know-about-digital-sat

Connolly, J. A., & Wantman, M. J. (1964). An exploration of oral reasoning processes in responding to objective test items. *Journal of Educational Measurement, 1*(1), 59–64.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*(4), 253–278.

Corbett, A. T., Anderson, J. R., & O'Brien, A. T. (1995). Student modeling in the ACT programming tutor. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment*, 19–41.

Csapo, B., Ainley, J., Bennett, R., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment (pp. 143–230). In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills*. Springer.

Deane, P. (2014). *Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks* (ETS Research Report No. RR-14-03). ETS. http://dx.doi.org/10.1002/ets2.12002

Deane, P., Roth, A., Litz, A., Goswami, V., Steck, F., Lewis, M., & Richter, T. (2018). *Behavioral differences between retyping, drafting, and editing: A writing process analysis* (ETS Research Memorandum No. RM-18-06). ETS.

Deane, P., Song, Y., van Rijn, P., O'Reilly, T., Fowles, M. E., Bennett, R. E., Sabatini, J. P., & Zhang, M. (2019). The case for scenario-based assessment of written argumentation. *Reading and Writing, 32*, 1575–1606.

DePascale, C., Dadey, N., & Lyons, S. (2016). *Score comparability across computerized assessment delivery devices*. Council of Chief State School Officers.

Desmarais, M. C., & Baker, R. S. J. d. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction, 22*(1–2), 9–38.

Dillon, G. F., & Clauser, B. E. (2009). Computer-delivered patient simulations in the United States Medical Licensing Examination (USMLE). *Simulation in Healthcare, 4*(1), 30.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions, pp. 1–33. https://doi.org/10.1145/3561048

Doignon, J.-P., & Falmagne, J.-C. (1999). *Knowledge spaces*. Springer-Verlag.

Dorans, N. J. (2000). Scaling and equating. In H. Wainer (Ed.), *Computer adaptive testing* (2nd ed., pp. 135–158). Erlbaum.

Dorans, N. J., & Walker, M. E. (2007). Sizing up linkages. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds), *Linking and aligning score scales* (pp. 179–198). Springer.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–515). American Council on Education/Praeger.

Duque, M. (2017). *Is there a PARCC mode effect?* Center for Education Policy Research. http://sdp.cepr.harvard.edu/files/cepr-sdp/files/sdp-fellowship-capstone-parcc-mode-pdf

EdTech Strategies. (2015). *Pencils down: The shift to on-line & computer-based testing*. https://www.edtechstrategies.com/wp-content/uploads/2015/11/PencilsDownK-8_EdTech-StrategiesLLC.pdf

Elo, A. (1978). *The rating of chessplayers, past and present*. Arco Publishers.

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

ETS. (2021). *Best practices for constructed-response scoring*. https://www.ets.org/content/dam/ets-org/pdfs/about/cr_best_practices.pdf

ETS, Pearson, & Measured Progress. (2016). *Final technical report for 2015 administration* (pp. 141, 150). PARCC. https://www.state.nj.us/education/assessment/district/PARCCTechReport15.pdf

Falmagne, J.-C., Koppen, M., Villano, M., Doignon, J., & Johanessen, L. (1990). Introduction to knowledge spaces: How to build, test, and search them. *Psychological Review, 97*(2), 201–224.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*, 193–202.

Fu, J., & Feng, Y. (2018). *A comparison of score aggregation methods for unidimensional tests in different dimensions* (ETS Research Report No. RR-18-01). ETS.

Fu, J., Chung, S., &. Wise, M. (2013). *Statistical report of Fall 2009 CBALTM writing tests* (ETS Research Memorandum No. 13-01). ETS.

Fu, J., Wise, M., & Chung, S. (2012). *Statistical report of Fall 2009 CBAL™ reading tests* (ETS Research Report No. RR-12-12). ETS.

Futia, G., & Vetro, A. (2019). On the integration of knowledge graphs into deep learning models for more comprehensible AI—three challenges for future research. *Information 2020, 11*(2), 122. https://doi.org/10.3390/info11020122

Gallagher, A., Bennett, R. E., Cahalan, C., & Rock, D. A. (2002). Validity and fairness in technology-based assessment: Detecting construct-irrelevant variance in an open-ended, computerized mathematics task. *Educational Assessment, 8*(1), 27–41.

Gee, J. P., & Shaffer, D. W. (2010). Looking where the light is bad: Video games and the future of assessment. *Edge, 6*(1), 3–19. http://edgaps.org/gaps/wp-content/uploads/EDge-Light.pdf

Georgia Department of Education. (2018). *Georgia's application for the Innovative Assessment Demonstration Authority under section 1204 of the Elementary and Secondary Education Act (ESEA)*. https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Documents/Flexibility/Georgia_IADA_Application.pdf

Gierl, M. J., & Haladyna, T. M. (Eds.). (2013). *Automatic item generation: Theory and practice*. Routledge.

Gobert, J. D., Moussavi, R., Li, H., Sao Pedro, M., & Dickler, R. (2018). Real-time scaffolding of students' online data interpretation during inquiry with Inq-ITS using educational data mining. In M. Auer, A. Azad, A. Edwards, & T. de Jong (Eds.), *Cyber-physical laboratories in engineering and science education*. Springer. https://doi.org/10.1007/978-3-319-76935-6_8

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education, 91*, 92–105.

G20. (2019). *G20 AI principles, section 1, article 1.3*. (*Annex to G20 ministerial statement on trade and the digital economy*). https://trade.ec.europa.eu/doclib/docs/2019/june/tradoc_157920.pdf

Guo, H., Rios, J. A., Haberman, S., Liu, L. O., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education, 29*(3), 173–183.

Guo, H., Deane, P. D., van Rijn, P. W., Zhang, M., & Bennett, R. E. (2018). Modeling basic writing processes from keystroke logs. *Journal of Educational Measurement, 55*, 194–216. https://doi.org/epdf/10.1111/jedm.12172

Guo, H., Zhang, M., Deane, P., & Bennett, R. E. (2019). Writing process differences in subgroups reflected in keystroke logs. *Journal of Educational and Behavioral Statistics, 44*, 571–596. https://doi.org/10.3102/1076998619856590

Guo, H., Zhang, M., Deane, P., & Bennett, R. E. (2020). Effects of scenario-based assessment on students' writing processes. *Journal of Educational Data Mining, 12*(1), 19–45. https://doi.org/10.5281/zenodo.3911797

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics, 40, 254–273.*

Hao, J., & Mislevy, R. J. (2018). *The evidence trace file: A data structure for virtual performance assessments informed by data analytics and evidence-centered design* (Research Report 18–28). ETS.

Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice, 43(2)*, 16–29. https://doi.org/10.1111/emip.12602

Haskins, C. (2019). Gaggle knows everything about teens and kids in school. *Buzzfeed*. https://www.buzzfeednews.com/article/carolinehaskins1/gaggle-school-surveillance-technology-education

Hattie, J., & Timperly, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112.

Heffernan, N., & Heffernan, C. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education, 24*(4), 470–497.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel* (MET Project Research Paper). Bill and Melinda Gates Foundation. https://danielsongroup.org/sites/default/files/inline-files/Reliabiilty_Observations_School_Personnel_Gates.pdf

Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp.5–30). Springer.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Praeger Publishers.

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment, 5*(2), 3–49.

Howard, R. A. (1960). *Dynamic programming and Markov processes*. Technology Press–Wiley.

Iacus, S. M., King, G., & Porro, G. (2011). *Causal inference without balanced checking: Coarsened exact matching*. Political Analysis. http://j.mp/iUUwyH

Institute for Education Sciences. (n.d.). *Progress in International Reading Literacy Study (PIRLS): Countries*. https://nces.ed.gov/surveys/pirls/countries.asp

Institute for Education Sciences. (2018). *Writing assessment*. https://nces.ed.gov/nationsreportcard/writing/

Irvine, S. H., & Kyllonen, P. C. (2010). *Item generation for test development*. Routledge.

Jewsbury, P. A., Finnegan, R., Xi, N., Jia, Y., Rust, K., & Burg, S. (2020). *2017 NAEP transition to digitally based assessments in mathematics and reading at Grades 4 and 8: Mode evaluation study*. National Center for Education Statistics. https://nces.ed.gov/nationsreportcard/subject/publications/main2020/pdf/transitional_whitepaper.pdf

Johns, J., & Woolf, B. (2006). A dynamic mixture model to detect student motivation and proficiency. *Proceedings of the National Conference on Artificial Intelligence, 21*, 163.

Johnson, S. (2019). *Students in California log on in record numbers to take online state tests*. EdSource. https://edsource.org/2019/students-in-california-log-on-in-record-numbers-to-take-online-state-tests/613321

Kapur, M. (2010). Productive failure in mathematical problem solving. *Instructional Science, 38*, 523–550.

Kingston, N. M (2009). Comparability of computer- and paper-administered multiple-choice tests for K–12 populations: A synthesis. *Applied Measurement in Education, 22*(1), 22–37.

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. J. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers and Education, 57*, 1813–1824.

Koedinger, K. R., & Corbett, A. (2006). Cognitive tutors: Technology bringing learning sciences to the classroom. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 61–77). Cambridge University Press.

Kolen, M.J. (1999). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment, 6*(2), 73–96. https://www.tandfonline.com/doi/abs/10.1207/S15326977EA0602_01

Kolen, M. (2007). Data collection designs and linking procedures. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp.31–55). Springer.

Kolen, M. J., & Brennan, R. L. (2004). *Testing, equating, scaling, and linking: Methods and practices*. Springer.

Kuang, C. (2017). Can A.I. be taught to explain itself? *New York Times Magazine*. https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-toexplain-itself.html

LaMar, M. M. (2018). Markov decision process measurement model. *Psychometrika, 83*(1), 67–88.

LaMar, M., Baker, R. S., & Greiff, S. (2017). Methods for assessing inquiry: Machine-learned and theoretical. *Design Recommendations for Intelligent Tutoring System, 5*, 137.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Lee, Y., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessment in Education, 2, 8*. https://doi.org/10.1186/s40536-014-0008-1

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*, 83–102.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement Issues and Practice, 13*(1), 5–8. https://doi.org/10.1111/j.1745-3992.1994.tb00778.x

Liu, J., Brown, T., Chen, J., Ali, U., Hou, L., & Costanzo, K. (2016). *Mode comparability study based on spring 2015 operational test data*. ETS.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing and guidance* (pp. 139–183). Harper & Row.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.

Lottridge, S. M., Nicewander, A. W., & Mitzel, H. C. (2010). Summary of online comparability studies for one state's end-of-course program. In P. Winter (2010), *Evaluating the comparability of scores from achievement test variations*. Council of Chief State School Officers (pp.13-32). http://www.ccsso.org/Documents/2010/Evaluating_the_Comparability_of_Scores_2010.pdf.

Madiega, T. (2019). *EU guidelines on ethics in artificial intelligence: Context and implementation* (PE 640.163). European Parliamentary Research Service. https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf

Maglieri, G., & Comande, G. (2017). Why a right to legibility of automated decision-making exists in the General Data Protection Regulation. *International Data Privacy Law, 7*, 243–265. https://doi.org/10.1093/idpl/ipx019

Margolis, M. J., & Clauser, B. J. (2006). A regression-based procedure for automated scoring of complex medical performance assessment. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 123–167). Erlbaum.

Marianti, S., Fox, J-P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics, 39*(6), 426–451. https://doi.org/10.3102/1076998614559412

Mills, C. N. (1999). Development and introduction of a computer adaptive Graduate Record Examinations General Test. In F. Drasgow & J. B. Olsen-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 117–135). Erlbaum.

Milne, R. B., Kelly, S. A., & Webb, D. C. (n.d.). *Effect of adaptivity on learning outcomes in an online intervention for rational number tutoring, "Woot Math," for Grades 3–6: A multi-site randomized controlled trial*. Saga Education, Inc. https://www.wootmath.com/efficacy

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods and prospects*. ETS.

Mislevy, R. J., & DiCerbo, K. E. (2012). Evidence centered design for learning and assessment in the digital world. In M. C. Mayrath, J. Clarke-Midura, D. H. Robinson, & G. Schraw (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (p. 13). Information Age.

Mislevy, R. J., & Gitomer, D. H. (1995). The role of probability-based inference in an intelligent tutoring system. *User Modeling and User-Adapted Interaction, 5*, 253–282 https://doi.org/10.1007/BF01126112

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (1999). A cognitive task analysis, with implications for designing a simulation-based assessment system. *Computers and Human Behavior, 15*, 335–374.

Mislevy, R. J., & Zwick, R. (2012). Scaling, linking, and reporting in a periodic assessment system. *Journal of Educational Measurement, 49*, 148–166.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2(2)*. https://doi.org/10.1016/j.rmal.2023.100050

Mullis, I. V. S. (2019). *eTIMSS: The future of TIMSS*. Boston College. http://timss2019.org/frameworks/framework-chapters/introduction/etimss-the-future-of-timss/

Mullis, I. V. S., & Prendergast, C. O. (2017). Developing the PIRLS 2016 achievement items. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016* (pp. 1.1–1.29). IEA.

National Center for Education Statistics. (2012). *Writing 2011: National Assessment of Educational Progress at Grades 8 and 12* (NCES 2021-470). U.S. Department of Education. https://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf

National Center for Education Statistics. (n.d.). *2014 Technology and Engineering Literary (TEL)*. U.S. Department of Education. https://www.nationsreportcard.gov/tel_2014/

Northwest Evaluation Association. (2019). *NWEA announces through-year assessment, a first-of-its-kind solution that blends proficiency and growth measurement*. https://www.nwea.org/content/uploads/2019/10/NWEA_Through-Year_Assessment_PR_OCT2019.pdf

Olson, L. (2019). *The new testing landscape: How state assessments are changing under the new federal Every Student Succeeds Act*. FutureEd. https://www.future-ed.org/wp-content/uploads/2019/09/REPORT_NewTestingLandscape-1.pdf

Oranje, A., Mazzeo, J., Xu, X., & Kulick, E. (2014). A multistage testing approach to group score assessments. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multi-stage testing: Theory and applications* (pp. 371–390). Chapman & Hall/CRC.

O'Reilly, T., Sabatini, J. P., & Wang, Z. (2019). Using scenario-based assessments to measure deep learning. In K. Millis, D. L. Long, J. P. Magliano, & K. Wiemer (Eds.), *Deep comprehension: Multi-disciplinary approaches to understanding, enhancing, and measuring comprehension* (pp.197–208). Routledge.

Organisation for Economic Co-operation and Development. (2017). *PISA 2015 technical report*. https://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf

Organisation for Economic Co-operation and Development. (2018). *PISA 2015 test questions*. https://www.oecd.org/pisa/test/pisa2015/#d.en.537240.

Organisation for Economic Co-operation and Development. (2019). *Recommendation of the Council on Artificial Intelligence, section 1, article IV, 1.3*. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449#_ga=2.100756866.2085259285.1559316172-1236900936.1559134188

Padilla, J.-L., & Benitez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*(1), 136–144.

Paek, P. (2005). *Recent trends in comparability studies*. Pearson Educational Measurement.

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2013). *Effectiveness of Cognitive Tutor algebra I at scale* (WR-984-DEIES). Rand Corporation. http://www.siia.net/visionk20/files/Effectiveness%20of%20Cognitive%20Tutor%20Algebra%20I%20at%20Scale.pdf

Pavlik, P. I., Jr., Cen, H., & Koedinger, K. R. (2009). Performance factors analysis—A new alternative to knowledge tracing. *Frontiers in Artificial Intelligence and Applications, 200* (1), 531-538. https://doi.org/10.3233/978-1-60750-028-5-531

Pearson. (2017, January 10). *Final technical report for the 2016 administration*. Illinois State Board of Education.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.

Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice, 30*(3), 10–28.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., & Sohl-Dickstein, J. (2015, December). Deep knowledge tracing. In *Proceedings of the 28th International Conference on Neural Information Processing Systems* (Vol. 1, pp. 505–513).

Pommerich, M. (2007). Concordance: The good, the bad and the ugly. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning score scales* (pp. 200–216). Springer.

Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating automated essay scoring: A (modest) refinement of the "gold standard." *Applied Measurement in Education, 28*(2), 130–142. https://doi.org/10.1080/08957347.2014.1002920

Pressey, S. L. (1926). A simple apparatus which gives tests and scores—and teaches. *School and Society, 23*(586), 373–376.

Pressey, S. L. (1927). A machine for automatic teaching of drill material. *School and Society, 25*(645), 549–552.

Protalinski, E. (2020). *How Duolingo uses AI in every part of its app*. Venture Beat. https://venturebeat.com/2020/08/18/how-duolingo-uses-ai-in-every-part-of-its-app/

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using item response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice, 35*(1), 38–47.

Ramineni, C., & Williamson, D. (2018), Understanding mean score differences between the *e-rater*® automated scoring engine and humans for demographically based groups in the *GRE*® General Test (Research Report 18-12). ETS. https://doi.org/10.1002/ets2.12192

Randall, J. Sireci, S., Li, X., & Kaira, L. (2012). Evaluating the comparability of paper- and computer-based science tests across sex and SES groups. *Educational Measurement: Issues and Practice, 31*(4), 2–12.

Rijmen, F. (2009). *Horizontal and vertical linking in a longitudinal design* (Research Memorandum 09-03). ETS.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review, 14*, 249–255.

Rivas, A., Fraile, J. M., Chamoso, P., González-Briones, A., Rodríguez, S., & Corchado, J. M. (2019). Students' performance analysis based on machine learning techniques. In L. Uden, D. Liberona, G. Sanchez, & S. Rodríguez-González (Eds.), *Learning technology for education challenges* (LTEC 2019, Vol. 1011, pp.428–438). Springer.

Rosenbaum, P. R., & Rubin, D. R. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33–38.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Russell, M. (2016). A framework for examining the utility of technology-enhanced items. *Journal of Applied Testing Technology, 17*(1), 20–32.

Russell, M., & Moncaleano, S. (2019). Examining the use and construct fidelity of technology-enhanced items employed by K–12 testing programs. *Educational Assessment, 24*(4), 286–304.

Russell, M., & Tao, W. (2004a). Effects of handwriting and computer-print on composition scores: A follow-up to Powers et al. Practical Assessment*, Research and Evaluation, 9*(1), 1–9. http://pareonline.net/getvn.asp?v=9&n=1

Russell, M., & Tao, W. (2004b). The influence of computer-print on rater scores. *Practical Assessment, Research and Evaluation, 9*(1), 1–13. http://pareonline.net/getvn.asp?v=9&n=10

Sandene, B., Horkay, N., Bennett, R. E., Allen, N., Braswell, J., Kaplan, B., & Oranje, A. (Eds.). (2005). *Online assessment in mathematics and writing: Reports from the NAEP Technology-Based Assessment Project* (NCES 2005-457). National Center for Education Statistics, U.S. Department of Education. http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005457

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerize adaptive testing: From inquiry to operation*. American Psychological Association.

Sawyer, R. (2007). Some further thoughts on concordance. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds), *Linking and aligning score scales* (pp. 217–230). Springer.

Schaeffer, G. A., Bridgeman, B., Golub-Smith, M. L., Lewis, C., Potenza, M. T., & Steffen, M. (1998). *Comparability of computer-adaptive and paper-and-pencil test scores on the GRE General Test* (Research Report 98-38). ETS.

Schaeffer, G. A., Steffen, M., Golub-Smith, M. L., Mill, C. N., & Durso, R. L. (1995). *The introduction and comparability of the GRE Computer-Adaptive General Test* (Research Report 95-20). ETS.

Settles, B., & Meeder, B. (2016). A trainable spaced repetition model for language learning. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 1848–1858). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215–232. http://www.jstor.org/stable/1435044

Shepard, L., Davidson, K., & Bowman, R. (2011). *How middle-school mathematics teachers use interim and benchmark assessment data* (CRESST Report 807). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153–189.

Shute, V. J., & Zapata-Rivera, D. (2010). Intelligent systems. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed., pp. 75–80). Elsevier.

Sireci, S. G. (2005). Unlabeling the disabled: A perspective on flagging scores from accommodated test administrations. *Educational Researcher, 34*(1), 3–12.

Sireci, S. G. (2020). Standardization and UNDERSTANDardization in educational assessment. *Educational Measurement: Issues and Practice, 39*(3), 100–105.

Sireci, S. G., & O'Riordan, M. (2020). Comparability issues in assessing individuals with disabilities. In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability issues in large-scale assessment* (pp. 178–204). National Academy of Education. https://doi.org/10.31094/2020/1

Sireci, S. G., Scarpati, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*, 457–490.

Skinner, B. F. (1958). Teaching machines. *Science, 128*(3330), 969–977.

Sleeman, D., & Brown, J. S. (1982). *Intelligent tutoring systems*. Academic Press.

Smarter Balanced Assessment Consortium. (2023). *Usability, accessibility, and accommodations guidelines*. The Regents of the University of California.

Smarter Balanced Assessment Consortium. (2024). *2024–2025 interim assessments overview*. The Regents of the University of California. https://portal.smarterbalanced.org/library/en/interim-assessments-overview.pdf

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 151–166.

Suppes, P. (1972). Computer-assisted instruction at Stanford [Conference proceedings]. In M. Marois (Ed.), *Man and computer* (pp. 298–330). Karger. https://suppes-corpus.stanford.edu/sites/g/files/sbiybj7316/f/computer-assisted_instruction_at_stanford_113.pdf

Tucker, B. (2012). Grand test auto: The end of testing. *Washington Monthly*. http://www.washingtonmonthly.com/magazine/mayjune_2012/special_report/grand_test_auto037192.php

Turek, M. (n. d.). *Explainable AI (XAI)*. Defense Advanced Research Projects Agency. https://www.darpa.mil/program/explainable-artificial-intelligence

United States Medical Licensing Examination. (2018). *Sample test questions: Step 3*. https://www.usmle.org/pdfs/step-3/Step3_Sample_Items.pdf

U.S. Department of Education. (n.d.). *Race to the Top assessment program*. https://www2.ed.gov/programs/racetothetop-assessment/index.html

U.S. Department of Education. (2012). *Writing 2011: National Assessment of Educational Progress at grades 8 and 12* (NCES-2012-470).

U.S. Department of Education. (2018). *A state's guide to the U.S. Department of Education's assessment peer review process*.

van der Linden, W. J. (2017). Test speededness and time limits. In *Handbook of item response theory* (pp. 249–266). Chapman and Hall/CRC.

van der Linden, W. J., & Lewis, C. (2015). Bayesian checks on cheating on tests. *Psychometrika, 80*(3), 689–706.

Vandermeulen, N., Leijten, M., & Van Waes, L. (2020). Reporting writing process feedback in the classroom: Using keystroke logging data to reflect on writing processes. *Journal of Writing Research, 12*(1), 109–140. https://doi.org/10.17239/jowr-2020.12.01.05

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., & Wintersgill, M. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence and Education, 15*(3), 147–204. http://www.andestutor.org/Pages/AndesLessonsLearnedForWeb.pdf

VanLehn, K., & Martin, J. (1998). Evaluation of an assessment system based on Bayesian student modeling. *International Journal of Artificial Intelligence in Education, 8*(2), 179–221.

Van Meer, E. (2003). Plato: From computer-based education to corporate social responsibility. *Iterations: An Interdisciplinary Journal of Software History, 2(1)*, 1–22. http://www.cbi.umn.edu/iterations/vanmeer.pdf

Veldkamp, B. P., & Paap, M. C. S. (2017). Robust automated test assembly for testlet-based tests: An illustration with analytical reasoning items. *Frontiers in Education, 7*, 1–8. https://www.frontiersin.org/articles/10.3389/feduc.2017.00063/full

von Davier, M., Khorramdel, L., He, Q., Shin, H. J., & Chen, H. (2019). Developments in psychometric population models for technology-based large-scale assessments: An overview of challenges and opportunities. *Journal of Educational and Behavioral Statistics, 44*(6), 671–705.

Wainer, H., Bradlow, E., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & Blas, C. A. W. (Eds.), *Computerized Adaptive Testing: Theory and Practice*, 245–269.

Wainer, H., & Eignor, D. R. (2000). Caveats, pitfalls, and unintended consequences of implementing large scale computerized testing. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 271–299). Erlbaum.

Wan, L. & Henly, G.A. (2012). Measurement properties of two innovative item formats in a computer-based test. *Applied Measurement in Education, 25*(1), 58–78.

Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement, 38*(1), 19–49.

Ward, W. C. (1988). The College Board computerized placement tests: An application of computerized adaptive testing. *Machine-Mediated Learning, 2*, 271–282.

Ward, W. C., Kline, R. G., & Flaugher, J. (1986). *College Board computerized placement tests: Validation of an adaptive test of basic skills* (Research Report 86-29). ETS.

Way, D., & Strain-Seymour, E. (2021). A framework for considering device and interface features that may affect student performance on the National Assessment of Educational Progress. American Institutes for Research. https://www.air.org/sites/default/files/Framework-for-Considering-Device-and-Interface-Features-NAEP-NVS-Panel-March-2021.pdf

Way, W. D., Davis, L. L., & Fitzpatrick, S. (2006, April 8–10). *Score comparability of online and paper administration of the Texas Assessment of Knowledge and Skills* [Paper presentation]. National Council on Measurement in Education Annual Meeting, San Francisco, CA, United States.

Way, W. D., Lin, C., & Kong, J. (2008, March 25–27). *Maintaining score equivalence as tests transition online: Issues, approaches and trends* [Paper presentation]. National Council on Measurement in Education Annual Meeting, New York, NY, United States.

Way, W. D., & Robin, F. (2016). The history of computer-based testing. In C. S Wells & M. F. Boyd (Eds.), *Educational measurement: From foundations to future* (pp. 185–207). The Guilford Press.

Way, W. D., Um, K., Lin, C., & McClarty, K. L. (2007, April 10–12). *An evaluation of a matched samples method for assessing the equivalence of online and paper test performance* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Chicago, IL, United States.

Weiss, D. J. (1976). Adaptive testing research at Minnesota: Overview, recent results, and future directions. In C. L. Clark (Ed.), *Proceedings of the first conference on computerized adaptive testing* (pp. 24–35). U.S. Civil Service Commission.

Weiss, D. J. (1978). *Proceedings of the 1977 computerized adaptive testing conference*. University of Minnesota.

Weiss, D. J. (1980). *Proceedings of the 1979 computerized adaptive testing conference*. University of Minnesota.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473–492.

Weiss, D. J. (1985). *Proceedings of the 1982 computerized adaptive testing conference*. University of Minnesota.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Erlbaum.

Winter, P. (2010). *Evaluating the comparability of scores from achievement test variations*. Council of Chief State School Officers. www.ccsso.org/Documents/2010/Evaluating_the_Comparability_of_Scores_2010.pdf

Wise, L. L. (2011). Picking up the pieces: Aggregating results from through-course assessments. ETS. https://www.ets.org/Media/Research/pdf/TCSA_Symposium_Final_Paper_Wise.pdf

Yeung, C. K., & Yeung, D. Y. (2018, June). Addressing two problems in deep knowledge tracing via prediction-consistent regularization. In *Proceedings of the Fifth Annual*

ACM Conference on Learning at Scale (pp. 1–10). Association for Computing Machinery.

Yu, L., Livingston, S. A., Larkin, K. C., & Bonett, J. (2004). *Investigating differences in performance between computer-based and handwritten essays* (RR-04-18). ETS.

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*(4), 337–362.

Zhang, M., Bennett, R. E., Deane, P., & van Rijn, P. W. (2019). Are there gender differences in how students write their essays? An analysis of writing processes. *Educational Measurement: Issues and Practice, 38*(2), 14–26. https://doi.org/10.1111/emip.12249

Zhang, M., & Deane, P. (2015). *Process features in writing: Internal structure and incremental value over product features* (Research Report 15-27). ETS.

Zhang, M., Deane, P., van Rijn, P. W., & Bennett, R. E. (2019). Scenario-based assessments in writing: An experimental study. *Educational Assessment, 24*, 73–90. https://doi.org/10.1080/10627197.2018.1557515

Zhang, M., Zou, D., Wu, A. D., Deane, P., & Li, C. (2017). An investigation of writing processes employed in scenario-based assessment. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 321–339). Springer.

NOTES

1. Soft-max is an extension of the logistic function to multiple-category variables, for example, $p(x_i)=\frac{\exp x_i}{\sum_j^k x_j}$.
2. That USMLE uses both analysis of product and response process in computing scores raises the question of how automated scoring differs from response process analysis in the consequential testing context. As suggested in the section introduction, automated scoring is generally used to produce one or more quantities for input into computing a test score. Response process analysis, in contrast, has been more often directed at such purposes as providing validity evidence, identifying possible guessing or cheating behavior, describing how groups differ in their approaches to problem-solving, and suggesting how instruction might be redirected. Automated scoring is widely used in operational consequential assessment. Response process analysis is far less prominent.
3. The methods used here are like the machine learning methods described for the evaluation of response processes in the prior section.
4. Relatively few studies have looked at this issue, but those that have been conducted suggest the presence of substantively important differences in how algorithms operate across at least some demographic groups (Bridgeman et al., 2012; Ramineni & Williamson, 2018).

5. There are many examples of current assessments taking a more traditional mastery test form that are used to support instruction but are not of the type described here. Measures like the Smarter Balanced interim assessment blocks are linear tests constructed to measure mastery of a narrow competency (or set of related narrow competencies) (Smarter Balanced, 2019). Results are linked to instructional resources in the Smarter Balanced Tools for Teachers.
6. Note that the focus here is on learning interactions only. In contrast, some schools record virtually every student online interaction for purposes of identifying safety threats (Haskins, 2019).