Research Article

Validity Considerations in Complex Task Design

Richard J. Tannenbaum and Irvin R. Katz, *Educational Testing Service*

Structured Abstract

- **Background**: With the growth of digital technology has come a desire to measure learning—both formatively and summatively—via complex performance tasks. Such tasks mimic the types of learning environments (whether digital or non-digital) familiar to learners, which is part of their appeal. Designed well, performance tasks present learners with similar affordances and challenges that they face in a non-assessment context, potentially allowing for greater generalizability (Conn et al., 2020). Yet complex and digital performance tasks may present some unique challenges in accumulating the necessary design-based and interpretation-based evidence of validity. In this paper, we discuss some of the measurement challenges and validity considerations associated with complex and digital performance tasks.
- **Challenges of Performance Assessments**: One measurement challenge presented by complex and digital performance tasks is that of reliability. Because the number of such tasks that can reasonably be included in any one occasion of testing is fewer than what is typical with more traditional task types, internal estimates of reliability tend to be lower. Furthermore, because these task types tend to be heavily contextualized, attention needs to be devoted to assuring that the intended construct is being measured, and not overshadowed by the context, such that what is being measured has more to do with the context than the construct of primary interest. Another challenge is that of scoring. While some performance tasks may be amenable to automated scoring, as the complexity of the tasks increases, human scoring most often needs to be applied, as automated scoring technologies are still somewhat limited. This not only has cost and time

implications, but also may introduce a source of construct-irrelevant variance or unreliability—scoring bias. Although, even automated scoring algorithms are not immune to bias, as they are calibrated using human-scored responses. Other issues related to construct sampling, fairness, and logistics are addressed in the manuscript.

- **Design of Complex Performance Assessments: Validity Considerations**: Complex performance tasks should be designed to be straightforward: Learners should understand what is being measured, how to interact with the task, how to indicate responses to the task, and so on. Clarity of the task design is important so that learners' solution processes and responses do not reflect construct-irrelevant factors. Because of the complexity of the tasks, it is common for assessment designers to take an iterative refinement approach whereby early concepts are tried out in simpler form and at a small scale before refining (or wholesale redesigning), followed by further testing and refining, leading to a fully deployed application (e.g., Katz et al., 2004). For convenience, we have segmented the creation of digital performance tasks into three phases: early prototype, initial assessment, and deployed assessment. We discuss the validity considerations germane to each of these phases.
- **Conclusions**: In this paper, we touched on the benefits of complex and digital performance tasks and on the considerations that must be acknowledged and acted upon to support the reasonable use of these tasks. While issues related to "design validity" apply to all types of assessments, the issues would appear to be elevated when considering complex and digital performance assessments, because of the need to account for the features, affordances, and functions that need to be included in assessment tasks to assure their authenticity, meaningfulness, and positive learner experience (engagement with the tasks). And such assurances can only be realized through iterative task design, tryout, and revision, coupled with the appropriate and purposeful collection of evidence that the task development is on the intended track. This validation process (evidence gathering) proceeds through the eventual deployment of the assessment, whereby outcomes-based impact may be observed, and the consequences of operational use may be evaluated. As assessment technology continues to evolve and the interactions between the learner and the assessment task become more involved, we will need to consider new measurement and scoring models, new models of validity, and new approaches to validation.

Keywords: complex task design, design validity, digital performances, reliability evidence, validity evidence, Workplace English Communication (WEC), writing analytics

1.0 Introduction

With the growth of digital technology in the classroom has come a desire to measure learning—both formatively and summatively—via complex performance tasks. Such tasks mimic the types of learning environments (whether digital or non-digital) familiar to learners, which is part of their appeal. Designed well, performance tasks present learners with similar affordances and challenges that they face in a non-assessment context, potentially allowing a clear inference about real-world achievement based on achievement during an assessment (Conn et al., 2020).

Well-designed performance tasks also support learning and teacher instruction (Darling-Hammond, 2014), therefore supporting formative uses. Formative assessment, in its simplest meaning (cf. Andrade et al., 2019; Bennett, 2011; Lyon et al., 2019), is when information yielded from an assessment is used by teachers and learners to improve learner achievement. Formative assessment contrasts with summative assessment, for which the primary objective is to understand learners' current state of knowledge, understanding, or skill. Black and Wiliam (2018) summarize the distinction between the two uses thusly:

> Where the inferences relate to the status of the student, or about their future potential, then the assessment is functioning summatively. Where the inferences relate to the kinds of actions that would best help the student learn, then the assessment is functioning formatively. (p. 553)

Note that well-designed[1] performance tasks may also provide summative information.

Whether used for formative or summative purposes, performance assessments must elicit from learners—and provide to teachers, learners, and other relevant decision makers—the information needed to fulfill the assessment's intended purpose. Thus, a central measurement question for performance assessments is one of validity: Does performance on complex tasks allow us to infer with confidence what learners know, are able to do, or need to develop? Similarly, for assessment situations beyond the classroom, performance assessments and other complex tasks raise validity issues for learners other than students, such as job candidates and prospective teachers. While validity has a long history in the design of traditional assessments and their use (e.g., Cronbach, 1971; Kane, 2006, 2016; Messick, 1989), complex performance tasks, which are also becoming more digital, bring unique challenges.

In this paper, we discuss some of the validity issues associated with the design of complex performance assessments. We first introduce performance assessments and the validity-centered motivations that drive their use, focusing primarily on digital performance assessment. We also consider measurement challenges: the issues that potentially undermine the validity of inferences

[1] We will often use the phrase "well-designed performance task" in this document. Our intention is to be clear that simply presenting a performance task to learners is not enough to assure the correct evidence of learner knowledge, skills, or abilities is being elicited. It is quite difficult to craft a performance task that achieves all of its goals, as we outline in this paper, so our statements about the benefits of performance assessments only apply to those that have been designed appropriately (i.e., "well designed").

one can draw about learners based on their performance. We then more formally introduce validity and validation—the goal of gathering logical and empirical support for the appropriateness of inferences made about performance on complex tasks. Based on an idealized three-phase process for the design of digital performance assessments, we illustrate the validity concerns that come into play at each phase. We conclude with a brief discussion of implications of well-designed digital performance tasks for the future of assessment.

2.0 Performance Assessments

Through performance tasks, learners demonstrate their knowledge and skills in a domain by creating a response (sometimes involving extended activities) rather than selecting from a set of responses as in traditional multiple-choice items. Examples include engaging with "standardized patients" (Ark et al., 2014), combining electronic components to design a circuit (Katz & James, 1998), conducting experiments to identify and explain the relationship between properties of paper towels and the amount of water absorbed (Shavelson et al., 1992), or demonstrating language proficiency through task-based language assessment (Norris, 2016). As an example of large-scale performance assessment, the National Assessment of Educational Progress (NAEP) began in the 1990s to supplement the traditional multiple-choice and constructed-response assessment by using "hands-on tasks" in which learners worked with physical materials to demonstrate scientific concepts and processes (National Research Council, 1996). Learners might be asked to measure the length of pencils that had been soaked in different solutions and to draw conclusions about the salt content of the solutions. As they solved the tasks, learners filled in worksheets that recorded elements of their solution, such as their measurements, hypotheses, and conclusions. Learners were scored based on the answers supplied in these worksheets (O'Sullivan et al., 1997). Portfolio-based teacher assessment, which requires individuals to compile support for their teaching competencies (e.g., video recordings of their teaching along with critical analysis) has been in place for years, and, although not without controversy, is still in practice (Whittaker et al., 2018). Similarly, the use of standardized patients, as noted above, requiring individuals to interact with actors portraying patients with various symptoms, remains a staple of healthcare preparation and assessment (Peters, 2019). Rupp et al. (2012) noted that the Cisco Networking Academy makes use of a digital platform for designing, administering, and scoring complex tasks related to the training and development of network engineers that makes use of representations of real-life equipment.

Continued advances in machine learning and data science have accelerated the application of learning analytics to assessment practices. Palmquist (2019) noted that learning analytics often addresses (1) estimating the probability of student success in a course of study, (2) identifying where and when teacher instruction and engagement may maximally support student learning, and (3) evaluating the effectiveness of instructional materials, interventions, and teaching practices. Analytic technologies also serve as a foundation for personalized learning; one recent example in writing is that of Writing Mentor^{TM}. Currently available as a Google Docs add-on, it provides automated feedback to learners on four major aspects of academic writing: credibility of claims, topic development, coherence, and editing (Burstein et al., 2018). Lastly, game-based assessment, a form of technology-rich assessment, offers advantages that more conventional forms of assessment do not. DiCerbo (2017) noted that game-based (or game-like) measures take advantage of learners' familiarity with and affinity for digital games, and so when properly designed, engage learners; and they present learners with activities in simulated real-world contexts, increasing the generalizability of inferences beyond the assessment. Game-based technologies also lend themselves to so-called stealth assessments, whereby assessments are seamlessly embedded in digital games, and so learners are unaware they are being assessed, as they acquire intended knowledge and skills from playing the games (Shute et al., 2009; Wang et al., 2015).

Tasks and technologies such as those mentioned above demonstrate how, in performance assessments, a learner "communicates his or her understanding of content, of process and strategy, and of the results obtained" (Baker & O'Neill, 1996, p. 186). In fact, often the scoring of performance assessments incorporates the procedures used to solve the problem or to demonstrate the competency. Learner procedures might be inferred from what they produce in the task, such as by scoring the equations manipulated in an algebra problem or the rationale behind an experimental design, or through more direct observation. Computer-delivered tasks are especially informative as they allow the possibility of scores based on the event logs (e.g., button presses, mouse clicks, pauses) generated as learners interact with the task (Ercikan & Pellegrino, 2017), also called "process data."

Many of the perceived benefits of performance assessments center on evidence: the evidence they elicit regarding learners' levels of proficiency. Performance assessments provide richer information about what a learner knows and can do compared with multiple-choice items or other forms of selected-response items (Stosich et al., 2018). For example, policymakers and researchers have argued that stricter goals for education and higher standards require the type of information available through performance assessment, such as evidence of complex and deep thinking (Baker, 1998; Darling-Hammond, 2017; Shavelson et al., 1992; Wiggins, 1993). Evidence of such complexity of thought might be elicited through learners' responses as well as through additional reflective prompts (Clark, 2012, discusses the role of self-reflection within formative assessment). Furthermore, compared with multiple-choice assessments, well-designed performance assessments may align more closely with ideas about better teaching practices (Baker et al., 1993; Ball et al., 2008; Darling-Hammond, 2014). As a result, they may encourage appropriate instruction (i.e., teaching "to the test" is the same as good teaching) and use (e.g., in a classroom to provide "on the ground" information to help guide teachers' instruction). Much of the accepted value of performance assessment may find its roots in employment (personnel) testing (Guion, 1965), in which work samples are often part of the assessment process. Work sample or performance testing involves a situation in which the person being tested performs practical tasks drawn from or based on the job itself (Siegel, 1986). In essence, work samples "are hands-on performance tests or simulations of the job" (Rogelberg, 2007, p. 905), increasing the fidelity of the knowledge, skills, and abilities (i.e., competencies) measured by the assessment to the competencies one is trying to generalize to outside of the assessment. The use of digital technology in the classroom has greatly extended the range of learning opportunities for learners, and the complexity of these tasks necessitates that learners be assessed in ways that align with, or have higher fidelity to, the ways they are taught (Quellmalz & Pellegrino, 2009).

3.0 Challenges of Performance Assessments

Along with the above benefits of performance tasks come several challenges, which are discussed in the following sections.

3.1 Reliability

Performance tasks tend to be less reliable than more discrete forms of measurement (Stecher, 2010). The lower reliability is partially a result of the length of performance tasks: within a set testing time, fewer performance tasks can be administered compared with short, discrete items, each of which would provide a distinct score. The complexity of performance tasks also leads to variability in performance. Shavelson and colleagues (1993) demonstrated that the measurement error associated with scores on performance tasks is primarily due to high variability between tasks, even if those tasks were designed to measure similar constructs. A similar outcome is often observed with the use of "assessment centers," whereby the exercises (tasks) used to measure the constructs (dimensions), even similar constructs across different tasks, tend to account for more of the performance variance (Lee et al., 2017).

Furthermore, even on the same tasks, learners display variability in their performance on different occasions. Shavelson et al. (1993) noted that many tasks may be needed to achieve sufficiently generalizable measurement. The National Assessment of Educational Progress addresses this limitation by taking a matrix sampling approach, achieving reliable measurement in the aggregate by administering different tasks to different learners (Stecher, 2010).

3.2 Scoring

By design, performance assessments elicit rich, complex performance from learners. This richness creates challenges for reliable human scoring—and is often beyond current capabilities of automated scoring—requiring complex scoring rubrics, systematic training procedures (including the use of annotated benchmarks or exemplars), and regular calibration of human scorers to maintain their consistent understanding, internalization, and application of the scoring rubrics. Because of the cost of human scoring, there are limits to the amount of data that can be gleaned from performance assessments: Each feature of performance to be scored adds cognitive load and time. This may lead to a tendency for performance assessments to be scored holistically, rather than analytically, which lessens the difficulty of scoring, but results in less information documented about the nature or quality of the performance and of the specific decision criteria applied by the raters (Harsch & Martin, 2013), limiting usefulness for certain purposes, such as formative applications or professional development.

There may be ways to reduce the costs of human scoring, such as by assembling teachers within the same school to score learners' responses. In addition to cost savings, such an approach may provide teachers with professional development opportunities as they acquire deep understanding of the scoring rubrics, are exposed to a large array of responses, and engage in peer discussions during scorer training and calibration. However, as digital collection of performance data becomes more widespread, another approach is to expand automated scoring of responses, using sophisticated scoring engines that can handle complex performance tasks, while also offering tailored and actionable feedback to teachers and learners.

3.3 Construct Sampling

Any set of assessment items or tasks necessarily samples from the universe of assessment tasks that could be placed before a learner. Because fewer evidence opportunities are placed before the learner, sampling of the task universe is important to avoid construct underrepresentation—the exclusion of aspects of the task universe that are significant to the assessment claim(s). Construct underrepresentation, therefore, is a source of irrelevant variance, and undermines inferences of validity (Nichols & Williams, 2009).

Construct representation is inclusive of the types of tasks one has a high probability of encountering in the non-assessment context. Such is the particular case of Workplace English Communication (WEC) termed Kitchen Design in this special issue of *The Journal of Writing Analytics*. As a form of WEC, Kitchen Design is a digitally delivered simulation involving complex tasks. Between 2018 and 2020, a prototype was developed using a scenario-based approach in which modules present opportunities for students to learn WEC by working in a fictitious company that specializes in designing and overseeing the construction of commercial and private kitchens. (For more about Workplace English Communication (WEC), see Oliveri et al., 2021, this issue; Corrigan & Slomp, 2021, this issue; and Slomp et al., 2021, this issue.) In the Kitchen Design assessment, domain representation is reflected in high-frequency occasions of workplace writing, such as conveying knowledge about the larger organization or the work team and clarifying assignments and project deliverables. Representation also includes that performance tasks are often contextualized to formalize a specific setting, set of circumstances, inter-dependencies, and cultural norms or expectations. A performance task, for example, that aims to address writing competence will likely need to be contextualized one way if the writing is to be academically focused, and another if it is to be workplace focused.

3.4 Logistics

Performance tasks tend to be costly to develop, to administer, and, as mentioned above, to score. Creating a performance assessment is not simply a matter of placing any real-world task in front of learners (Baker, 1998; Rupp et al., 2010). Tasks must be carefully selected to represent the domain (to avoid construct underrepresentation) and to elicit the behaviors that provide clear evidence of the intended learner knowledge, skills, or abilities (KSAs).

Performance tasks tend to be longer than discrete test items. However, when creating even a short, discrete test item, much expertise goes into crafting a question that elicits KSAs relevant to the domain of interest. If a learner can provide the correct answer without employing the expected KSAs or gives an incorrect answer but does possess the requisite KSAs, it calls into question the validity of a person's score. Short, discrete items, when developed with care, go through many reviews for clarity, fairness, language, and so on. Creating a longer, more involved performance assessment is both a more challenging task and requires more reviews, cognitive labs and usability studies, tryouts and field testing, and often greater information technology resources. From a testing company standpoint, a performance assessment represents a significant investment. While a discrete item might be eliminated from a test or replaced by another item if data suggests that it has poor statistical qualities, performance tasks cannot be so easily discarded or replaced because of the significant investment in their creation. That creation includes the performance task itself as well as scoring rubrics and training procedures for human raters. (Some types of performance tasks may be amenable to automated scoring, which, while improving scoring efficiency, also comes with non-trivial costs, largely due to the creation and fine-tuning [training and calibration] of the algorithms that undergird the automated scoring system.) All of this is not to say that performance assessments cannot be feasibility delivered at a large scale; after all, there are many examples of large-scale performance assessments (e.g., Advanced Placement® [AP] Art and Design, United States Medical Licensing Examination®, National Board for Professional Teaching Standards®; Stecher, 2010).

3.5 Fairness

The notion of fairness includes multiple aspects. Performance assessments may be considered fair because, if developed properly, they align more closely with settings, situations, circumstances, and cultural norms or expectations reflective of the targeted non-assessment context—the context to which we ultimately want to generalize—and so tap into relevant and important KSAs. A task that requires a teacher to facilitate a small-group discussion in a simulated classroom, for example, is fair, because that task is a close reflection of what a teacher needs to be able to do in actual classroom practice—a teacher needs to be able to engage learners in content-relevant discourse and learner-to-learner interaction. A flight simulator is fair, because pilots, when on the job, need to be able to land the plane safely. In general, a task that looks more like the actual non-test task of ultimate interest and that engages learners to demonstrate competencies of ultimate interest tends to be consider fair.

But that is only one aspect of fairness with regards to performance assessments. Another source of fairness was highlight above: scoring. A task or set of tasks may do a good job of eliciting evidence of relevant competencies, but the scoring of that evidence may not be as reliable or accurate as necessary. This depends on the intended use of the assigned scores (lower or higher stakes; see Tannenbaum & Kane, 2019), or there may be certain human scorers that constantly rate too low or too high (and so are a source of systematic bias). Of course, even automated scoring systems are not immune to systematic bias (Madnani et al., 2017). Automated scoring systems rely on human-scored performance samples to "learn and calibrate." As a result, although the system may be able to apply scoring in a more consistent manner than humans (machines do not fatigue or demonstrate drift), if they are trained and calibrated with biased pre-scored samples, the automated scoring system will reflect that built-in bias.

To achieve the promise of performance assessments while mitigating the challenges requires careful planning, designing, and evaluating. In other words, attention to comprehensive, integrative validity and validation is needed. Validity encompasses design, development, implementation, scoring, and reporting. Each stage in the assessment process should be backed by relevant and appropriate evidence.

4.0 Validity and Validation: What They Are and Why They Are Important for Complex Task Design

One uses the information produced from an assessment, whether that assessment is performance based, digital, or more traditional, to draw conclusions about what learners know, can do, or need, and to inform actions or decisions based on the information. Validity relates to the logical and empirical support for the reasonableness and appropriateness of the conclusions drawn, actions taken, or decisions made.

More formally stated, "validity refers to the degree to which evidence[2] and theory support the interpretations of scores for proposed uses of tests" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014, p. 11). Validation refers to the methods, approaches, and processes used to collect and articulate support for validity (AERA, APA, & NCME, 2014). As many have noted, including Kane (2006, 2016), validity is not about the assessment itself, but about the inferences one makes from the assessment information (most often scores).

There are multiple sources of validity evidence, and depending on the nature and kind of inferences one wants to make from the assessment information, as well as the conditions and context within which the testing occurs, certain sources take on greater importance (Tannenbaum & Kane, 2019). For example, if the intended use is one of prediction, as in the case of an undergraduate admissions test, then evidence that performance on the assessment is related to undergraduate grades or other valued undergraduate outcomes is important. If the intended use is

[2] The term "evidence" is used in both the literature on validity and in discussions of assessment design (most notably "evidence-centered design"). For validity, evidence refers to the empirical and logical support for the overall validity argument (e.g., Kane, 2006) of an assessment. For assessment design, evidence refers to the observable learner behaviors that provide support for the claims of the assessment (e.g., that a learner has particular knowledge or skills; e.g., Mislevy & Riconscente, 2006). For clarity, in this paper we mostly use the term "support" in the context of validity and "evidence" in the context of assessment design.

one of occupational licensure, then evidence is needed that the assessment content reflects KSAs important for professional practice and the passing score was set following recommended practices. Both admissions and licensure are higher stakes uses. On the other hand, formative assessment tends to be considered a lower stakes use for learners (Dixson & Worrell, 2016). Even so, evidence that the assessment measures the KSAs it was intended to measure and supports the level of learning it was intended to foster is needed. Formative assessment use does not pardon us from establishing a reasonable basis of validity. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) describe the different sources of validity evidence (test content, response processes, internal structure, relations to other variables, and consequences of testing) in detail.

Validity and validation are about accumulating evidence to support intended claims to be made from assessment information. But to believe that they only come into effect *after* an assessment has been developed and administered would be a misconception. Validity considerations must also come into play as an assessment is being conceptualized and constructed. Briggs (2004) refers to this as design validity, which he notes is closely related to Kane's (2006, 2016) interpretive argument. Central to design validity is a careful articulation of the theory and pathways that underlie the expected changes or outcomes to occur from the assessment information use (one may think of this as a theory of action). It is to this "blueprint" that all other facets of the assessment-construction process, for example, task design and development, scoring approaches, and reporting applications, are aligned. Having this blueprint up front makes it likely that the assessment will fulfill its intended purpose and use, because all subsequent development-related activities can be aligned to and vetted against the blueprint. Evidence-Centered Design (e.g., Mislevy & Riconscente, 2006) is a systematic approach to assessment development that serves to formalize alignment among the test purpose, desired claims, evidence and tasks, and, as such, provides a solid foundation for validity (Zieky, 2014).

This alignment perspective is important for all forms of assessment, but takes on greater importance for complex performance tasks, in part, because such tasks are contextualized, or framed around authentic situations or settings (Bachman & Palmer, 2010; Kane et al., 1999; Mislevy, 2016). Part of the design validity for such tasks, therefore, needs to account for the different factors and variables ("demand features") that define the setting or context of interest and moderate performance on that task. This not only places more of a burden on task design and development but is also likely to increase opportunities for inadvertently introducing sources of construct-irrelevant variance—threats to validity. Drawing validity evidence out of complex performance tasks involves much more than considerations of correlations of scores or factor analyses (Katz et al., 2017).

Thus, the design of complex performance tasks should focus on the evidence—the observable behaviors that indicate levels of the construct (knowledge, skills, or abilities) of interest. Specifying the construct, evidence, and nature of the tasks that will elicit useful evidence are central for the design of any assessment. For complex performance tasks, given their many interacting activities and the richness of the potential data to be collected, consistent documentation of the task features and how they relate to the construct is particularly important. Such articulation of the underlying argument of the tasks and how they fit together into the larger assessment form key elements of design validity that ultimately support intended inferences about learner competencies. Katz et al. (2017) provide illustrations of such interplay among types of validity evidence for complex, digital performance assessments. The researchers analyze the validation needs related to two game-based assessment approaches: conversation-based assessment (test-takers "converse" with simulated agents) and stealth assessments (discussed earlier in this paper). Validity evidence ranges from the traditional (quantitative relationship among game scores and academic measures) to the modeling of in-game behaviors (i.e., process data) against expected models of performances.

One construct that continues to garner attention is that of Workplace English Communication. English is recognized as the lingua franca of workplace communication (Nickerson, 2005), and less than acceptable workplace English may have systemic consequences. In the context of international business, Oliveri and Tannenbaum (2019) observed that "low English proficiency has a negative impact on various stakeholders including countries, businesses seeking to expand beyond national borders, employers seeking to run international businesses successfully and employees seeking greater job opportunities, higher salaries and promotions" (p. 345). In 2009, Ekkens and Winke raised concerns that standardized tests used to measure workplace English were not adequate, as they did not, for example, align closely with course objectives or curricula, and did not offer adequate measurement of very low English proficiency. Advances in both digital language training and evaluation would seem to lessen concerns raised by Ekkens and Winke. One such advancement is mobile-assisted language learning (MALL; Stockwell & Hubbard, 2013), that is, learning and assessment via smartphones and tablets, an extension of computer-assisted language learning. The advantages of MALL, according to Kukulska-Hulme and Viberg (2018), include the following: flexibility to learn at times convenient to the learner, personalized learning, timely feedback, peer coaching, self-evaluation, and cultural authenticity. They also note that MALL facilitates collaborative learning (CL), which includes practicing conversations in the target language. Clearly, however, technology in and of itself does not guarantee success in language learning or assessment; that success is still dependent on coupling that technology with appropriate pedagogical strategies, learning science principles, and using evidence-based assessment practices.

5.0 Design of Complex Performance Assessments: Validity Considerations

Complex performance tasks should be designed to be straightforward: Learners should understand what is being measured, how to interact with the task, how to indicate responses to the task, and so on. Clarity of the task design is important so that learners' solution processes and responses do not reflect construct-irrelevant factors. Because of the complexity of the tasks, it is common for assessment designers to take an iterative refinement approach whereby early concepts are tried out in simpler form and at a small scale before refining (or wholesale redesigning), followed by further testing and refining, leading to a fully deployed application (e.g., Katz et al., 2004). For convenience, we will segment the creation of digital performance tasks into three "phases": early prototype, initial assessment, and deployed assessment. We recognize that this characterization is incomplete. Nevertheless, these three phases are useful for the discussion of validity issues and validation strategies, as each phase involves distinct validity concerns and questions that should be addressed as well as validity evidence to be collected.

5.1 Early Prototype Stage

First, as noted above with respect to design validity, the creation of a complex performance assessment typically begins with an articulation of what information the assessment, overall, is intended to provide, and, specifically, what each task is therefore intended to elicit from learners—the connection from observable learner performance to interpretation of those responses to the claims about learners (e.g., achievement on particular knowledge and skills). This articulation is the basis for an initial design for a task, which might be realized as a low-fidelity prototype, such as a sketch, a storyboard, or a wireframe. In the parlance associated with lean startups, this early-stage prototype may be considered a "minimally viable product," functional enough to support small-scale data collection and user reactions (see Ries, 2011). This type of prototype supports collecting data via usability studies to test whether the basic assumptions about the task hold: Do learners understand what they are supposed to do? Are they engaging with the task as expected? Are there aspects of the task they misinterpret?

Sometimes a usability trial includes alternative versions of the design: different ways of achieving the same goals. This is especially relevant for complex performance tasks whereby the designers may include variations of task features or situational elements to see which versions function best or come closer to achieving what was intended. Often this occurs by randomly assigning intended learners to the different task variations (conditions) and collecting evidence about their reactions, engagement, and, preliminarily, their level of task performance. Based on the initial usability trials, the designers refine their design or redesign what was created.

This process of initial creation and preliminary evidence-based feedback proves useful to gauge how close the current task version is to eliciting the kind of response processes that were intended. If, for example, a task was intended to elicit an intermediate level of spoken language proficiency, but feedback from learners indicate that the task is not sufficiently linguistically challenging, that is important early evidence that the task is not on-track to provide valid and meaningful information. Of course, learners must also then be prompted to provide direct and constructive recommendations about how the task might be enhanced or improved. At this stage in the task-development process, design validity is central.

5.2 Initial Assessment Stage

Once the minimally viable product is constructed—over the course, no doubt, of several iterations of build and trial—a more fleshed out version of the task or set of tasks must be constructed, trialed at larger scale, and evaluated. But even at this stage, we are still not talking about an operational task, set of tasks, or intact performance assessment ready for deployment to learners in contexts with consequence. We are moving closer to that readiness, as this next version of the task is intended to function similarly to the eventual operational version.

The key questions at this stage are closer to typical assessment validity questions. For example, does the task address the construct(s) of interest and elicit the type of responses that were intended? For example, if a task is intended to tap critical thinking skills, does it, or does it only tap routine decision-making skills? Is the content and context of the task consistent with expert opinion about what's needed to assure relevance, authenticity, and generalizability? Answers to such questions will likely not yield the amount of data needed to do inferential statistical analyses but will provide important information from potential learners on how they understand, interpret, and engage with the task(s).

Sources of validity evidence at this stage may come, for example, from learners verbalizing how they approached and solved the task (cognitive laboratories or think-alouds, e.g., Leighton, 2017); from analyzing keystroke logs, for example, to understand writing strategies (e.g., Leijten & Van Waes, 2013); and/or from eye-tracking data to consider what aspects of the task description or task instructions learners devote more or less attention to (e.g., Keehner et al., 2016). Also important at this stage is working through several of the validity-related issues we raised earlier, such as scoring design issues (e.g., types of rubrics to use, level of rubric detail needed), rater training and calibration (assuming human scorers), and score reporting, as well as gathering indications of potential issues with adverse impact or differential performance by subgroups (e.g., racial/ethnic, gender, cultural, linguistic).

5.3 Deployed Assessment Stage

Once in operational use, and assuming sufficiently large and representative numbers of learners have engaged with the assessment, there is the opportunity to conduct statistical analyses that offer quantitative types of evidence. These would include reliability and/or generalizability, task difficulty and discrimination, differential task functioning, differential performance by subgroups and adverse impact, decision accuracy (if scores are used for placement or classification purposes), and additional validity-related evidence, such as analyses of the internal structure of the assessment—relationships among the different tasks and the stability of that structure across different groups of learners (Rios & Wells, 2014) and the relationship between performance on the assessment and other measures or outcomes of interest, such as the use of GRE® scores to predict grades in law school (Klieger et al., 2018). Validity evidence at this stage also extends to how decision-makers use the reported information and the consequences of that use. The expectation is that decision-makers will act on the reported information as intended. However, as

Tannenbaum (2019) notes, “unintended consequences may arise from misunderstanding the meaning of the scores, including attributing more meaning to the scores than justified” (p. 17).

6.0 Conclusions

Performance-based assessment is not new, but the value of these assessments to support formative use has become more obvious (Gorin & Mislevy, 2013) relatively recently. Advances in technology have opened the door to new ways of measuring skills and competencies that could not be adequately measured before, such as systems thinking, creative problem solving, and teamwork (Shute, 2011). Yet this opportunity also implies that greater care and attention must be placed on task design and development to assure, for example, fairness, relevance, and domain representation, and on the accumulation of evidence that supports the intended use of the assessment information. DiCerbo et al. (2017) reflected on some of the challenges presented by complex, digital assessments, including the need for measurement models to account both for responses across a sequence of assessment tasks, as well as the adaptive presentation of tasks or task sequences to learners based on their prior responses. They also reminded us of the potential fairness issues with digital assessments, where, even today, there is unequal access to technology among learners from different socioeconomic levels, leading to an experience gap with that technology and adding to observed performance differences. DiCerbo (2017), when commenting on the range of process-based constructs that can now be measured with digital assessments, also noted, however, “it is not clear how to take a log file of activity stream data and identify how the elements should be transformed into observable variables and included in measurement models” (p. 8). In addition, the fidelity to the real world afforded by complex, digital assessments also means that context (social and cultural) becomes a more important driver of response (and score) variance. Measurement models must continue to evolve to account for this source of variance as a main effect, but also how it may interact with different learner populations.

In this paper, we touched on the benefits of complex and digital performance tasks and on the considerations that must be acknowledged and acted upon to support the reasonable use of these tasks. While issues related to “design validity” apply to all types of assessments, the issues would appear to be elevated when considering complex and digital performance assessments, because of the need to account for the features, affordances, and functions that need to be included in assessment tasks to assure their authenticity, meaningfulness, and positive learner experience (engagement with the tasks). And such assurances can only be realized through iterative task design, tryout, and revision, coupled with the appropriate and purposeful collection of evidence that the task development is on the intended track. This validation process (evidence gathering) proceeds through the eventual deployment of the assessment, whereby outcomes-based impact may be observed, and the consequences of operational use may be evaluated. It is safe to predict that as assessment technology continues to evolve and the interactions between the learner and the assessment task become more involved, we will need to consider new measurement and scoring models, and new models of validity and new approaches to validation. We look forward to what the future state of assessment and evidence holds.

Author Biographies

Richard J. Tannenbaum, Ph.D., is an Associate Vice President in the Assessment and Learning Technology Research and Development Division of Educational Testing Service (ETS). Prior to this role, Richard was a Senior Director for the ETS Center for Validity Research, and a General Manager of Research. He has published numerous technical reports, journal articles, and book chapters on topics such as standard setting, validity, score reporting, and mapping test scores to language frameworks.

Irvin R. Katz, Ph.D., retired in 2020 as Senior Research Director of the Cognitive and Technology Sciences Center at Educational Testing Service (ETS). Throughout his 30-year career at ETS, he conducted research at the intersection of cognitive psychology, psychometrics, and technology. He is also a human-computer interaction practitioner with almost 40 years of experience in designing, building, and evaluating software for research, industry, and government.

References

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

Andrade, H. L., Bennett, R. E., & Cizek, G. J. (Eds.). (2019). *Handbook of formative assessment in the disciplines*. Routledge.

Ark, T. A., Ark, N., & Zumbo, B. D. (2014). Validation practices of the Objective Structured Clinical Examination (OSCE). In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 267-288). Springer.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Baker, E. L. (1998). *Model-based performance assessment* (CSE Technical Report #465). CRESST; UCLA.

Baker, E. L., & O'Neill, H. F., Jr. (1996). Performance assessment and equity. In M. B. Kane & R. Mitchell (Eds.), *Performance assessment: Problems, promises, and challenges* (pp. 183-199). American Institutes for Research.

Baker, E. L., O'Neill, H. F., Jr., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessments. *American Psychologist*, *48*, 1210-1218.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*, 389-407.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5-25.

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, *25*(6), 551-575.

Briggs, D. C. (2004). Comment: Making an argument for design validity before interpretive validity. *Measurement*, *2*, 171-174.

Burstein, J., Elliot, N., Beigman-Klebanov, B., Madnani, N., Napolitano, D., Schwartz, M., Houghton, P., & Molloy, H. (2018). Writing Mentor$^{\text{TM}}$: Writing progress using self-regulated writing support. *Journal of Writing Analytics*, *2*, 285-313.

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, *24*, 205-249.

Conn, C. A., Bohan, K. J., Pieper, S. L., & Musumeci, M. (2020). Validity inquiry process: Practical guidance for examining performance assessments and building a validity argument. *Studies in Educational Evaluation*, *65*, 1-11.

Corrigan, J. A., & Slomp, D. H. (2021). Articulating a sociocognitive construct of writing expertise for the digital age. *The Journal of Writing Analytic*s, *5*.

Cronbach, L. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). American Council on Education.

Darling-Hammond, L. (2014). Testing to, and beyond the Common Core. *Principal*, *93*(3), 8-12.

Darling-Hammond, L. (2017). *Developing and measuring higher order skills: Models for state performance assessment systems*. Council of Chief State School Officers.

DiCerbo, K. E. (2017). Building the evidentiary argument in game-based assessment. *Journal of Applied Testing Technology*, *18*, 7-18.

DiCerbo, K. E., Shute, V., & Kim, Y. J. (2017). The future of assessment in technology rich environments: Psychometric considerations of ongoing assessment**.** In J. M. Spector, B. Lockee, & M. Childress (Eds.), *Learning, design, and technology: An international compendium of theory, research, practice, and policy*. Springer.

Dixson, D. D., & Worrell, F. C. (2016). Formative and summative assessment in the classroom. *Theory Into Practice*, *55*(2), 153-159.

Ekkens, K., & Winke, P. (2009). Evaluating workplace English language programs. *Language Assessment Quarterly*, *6*, 265-287.

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

Gorin, J. S., & Mislevy, R. J. (2013). *Inherent measurement challenges in the next generation science standards for both formative and summative assessment* (K-12 Center at Educational Testing Service Invitational Research Symposium on Science Assessment). ETS.

Guion, R. M. (1965). *Personnel testing*. McGraw Hill.

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice*, *20*(3), 281-307.

Hoffman, B. J., Kennedy, C. L., LoPilato, A. C., Monahan, E. L., & Lance, C. E. (2015). A review of the content, criterion-related, and construct-related validity of assessment center exercises. *Journal of Applied Psychology*, *100*(4), 1143-1168.

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

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, *23*, 198-211.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement*: *Issues and Practice*, *18*(2), 5-17.

Katz, I. R., Williamson, D. M., Nadelman, H. L., Kirsch, I., Almond, R. G., Cooper, P. L., Redman, M. L., & Zapata-Rivera, D. (2004, June). *Assessing information and communications technology literacy for higher education* [Paper presentation]. Annual Meeting of the International Association for Educational Assessment, Philadelphia, PA, United States.

Katz, I. R., & James, C. M. (1998). Toward assessment of design skills in engineering (ETS Rep. No. RR-97-16). Educational Testing Service.

Katz, I. R., LaMar M. M., Spain, R., Zapata-Rivera, J. D., Baird, J.-A., & Greiff, S. (2017). Validity issues and concerns for technology-based performance assessment. In R. A. Sottilare, A. C. Graesser, X. Hu, & G. Goodwin (Eds.), *Design recommendations for intelligent tutoring systems* (Vol. 5, pp. 209-224). Army Research Laboratory.

Keehner, M., Gorin, J. S., Feng, G., & Katz, I. R. (2016). Developing and validating cognitive models in assessment. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 75-101). Wiley Blackwell Press.

Klieger, D. M., Bridgeman, B., Tannenbaum, R. J., Cline, F. A., & Olivera-Aguilar, M. (2018). *The validity of GRE® General Test scores for predicting academic performance at U.S. law schools* (Research Report No. RR-18-26). Educational Testing Service.

Kukulska-Hulme, A., & Viberg, O. (2018). Mobile collaborative language learning: State of the art. *British Journal of Educational Technology*, *49*, 207-218.

Lee, J., Connelly, B. S., Goff, M., & Hazucha, J. F. (2017). Are assessment center behaviors' meanings consistent across exercises? A measurement invariance approach. *International Journal of Selection and Assessment*, *25*(4)*,* 317-332.

Leighton, J. P. (2017). *Using think-aloud interviews and cognitive labs in educational research*. Oxford University Press.

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication*, *30*(3), 358-392.

Lyon, C. J., Nabors Oláh, L., & Wylie, E. C. (2019). Working toward integrated practice: Understanding the interaction among formative assessment strategies. *The Journal of Educational Research*, *112*(3), 301-314,

Madnani, N., Loukina, A., von Davier, A., Burstein, J., & Cahill, A. (2017). Building better open-source tools to support fairness in automated scoring. In *Proceedings of the First Workshop on Ethics in Natural Language Processing* (pp. 41-52). Association for Computational Linguistics.

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

Mislevy, R. J. (2016). How developments in psychology and technology challenge validity argumentation. *Journal of Educational Measurement*, *53*, 265-292.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61-90). Erlbaum.

National Research Council. (1996). *National science education standards.* National Academies Press.

Nichols, P. D., & Williams, N. (2009). Consequences of test score use as validity evidence: Roles and responsibilities. *Educational Measurement: Issues and Practice*, *28*(1), 3-9.

Nickerson, C. (2005). English as a lingua franca in international business contexts. *English for Specific Purposes*, *24*, 367-380.

Norris, J. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics*, *36*, 230-244.

Oliveri, M. E., Slomp, D. H., Elliot. N., Rupp, A. A., Mislevy, R. J., Vezzu, M., Tackitt, A., Nastal, J., Phelps, J., & Osborn, M. (2021). Introduction: Meeting the challenges of Workplace English Communication in the 21st century. *The Journal of Writing Analytic*s, *5*.

Oliveri, M. E., & Tannenbaum, R. J. (2019). Are we teaching and assessing the English skills needed to succeed in the global workplace? In V. M. Hammler & S. V. Palsole (Eds.), *The Wiley handbook of global workplace learning* (pp. 343-354). Wiley.

O'Sullivan, C. Y., Jerry, L., Ballator, N., & Herr, F. (1997). *NAEP 1996 science state report for Tennessee*. National Center for Education Statistics. http://nces.ed.gov/nationsreportcard/pdf/stt1996/97499tn.pdf

Palmquist, M. (2019). Directions in writing analytics: Some suggestions. *The Journal of Writing Analytics*, *3*, 1-12.

Peters, G. (2019). The role of standardized patient assessment forms in medical communication skills education. *Qualitative Research in Medicine & Healthcare*, *3*(2), 76-86.

Quellmalz, E. S., & Pellegrino, J. W. (2009). Technology and testing. *Science*, *323*(5910), 75-79.

Ries, E. (2011). *The lean startup*. Crown Business.

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema*, *26*(1), 108-116.

Rogelberg, S. G. (2007). *Encyclopedia of industrial and organizational psychology (Vol. 2),* Sage Publications.

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *Journal of Technology, Learning, and Assessment*, *8*(4).

Rupp, A. A., Levy, R., DiCerbo, K. E., Sweet, S. J., Crawford, A. V., Caliço, T., Benson, M., Fay, D., Kunze, K. L., Mislevy, R. J., & Behrens, J. T. (2012). Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. *JEDM | Journal of Educational Data Mining*, *4*(1), 49-110.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*(3), 215-232.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher*, *21*(4), 22-27.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503-524). Information Age Publishers.

Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Bitterfield, M. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295-321). Routledge; Taylor, & Francis.

Siegel, A. I. (1986). Performance tests. In R. A. Berk (Ed.), *Performance assessment: Methods & applications* (pp. 121-142). The Johns Hopkins Press.

Slomp, D. H., Oliveri, M. E., & Elliot. N. (2021). Afterword: Meeting the challenges of Workplace English Communication in the 21st Century. *The Journal of Writing Analytics*, *5*.

Stecher, B. (2010). *Performance assessment in an era of standards-based educational accountability*. Stanford University, Stanford Center for Opportunity Policy in Education.

Stockwell, G., & Hubbard, P. (2013). *Some emerging principles for mobile-assisted language learning*. The International Research Foundation for English Language Education. http://www.tirfonline.org/english-in-the-workforce/mobile-assisted-language-learning

Stosich, E. L., Snyder, J., & Wilczak, K. (2018). How do states integrate performance assessment in their systems of assessment? *Education Policy Analysis Archives*, *26*(13)*,* 1-31.

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

Tannenbaum, R. J., & Kane, M. T. (2019). *Stakes in testing: Not a simple dichotomy but a profile of consequences that guides needed evidence of measurement quality* (Research Report No. RR-19-19). Educational Testing Service.

Wang, L., Shute, V., & Moore, G. R. (2015). Lessons learned and best practices of stealth assessment. *Journal of Gaming and Computer-Mediated Simulations*, *7*, 66-87.

Whittaker, A., Pecheone, R., & Stansbury, K. (2018). Fulfilling our educative mission: A response to edTPA critique. *Educational Policy Analysis Archives*, *26*, 1-20.

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan*, *75*(3), 200-208, 210-214.

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa*, *20*(2), 79-87.