Chapter 1
The Validity of Technology Enhanced Assessments—Threats and Opportunities

Saskia Wools, Mark Molenaar and Dorien Hopster-den Otter

Abstract Increasing technological possibilities encourage test developers to modernize and improve computer-based assessments. However, from a validity perspective, these innovations might both strengthen and weaken the validity of test scores. In this theoretical chapter, the impact of technological advancements is discussed in the context of the argument-based approach to validity. It is concluded that the scoring and generalization inference are of major concern when using these innovative techniques. Also, the use of innovative assessment tasks, such as simulations, multi-media enhanced tasks or hybrid assessment tasks is quite double-edged from a validity point of view: it strengthens the extrapolation inference, but weakens the scoring, generalization and decision inference.

1.1 Introduction

Increasing technological possibilities encourage test developers to improve computer-based assessment in multiple ways. One example is the use of authentic items that have the potential to improve construct representation, making it possible to assess complex constructs like skills or competences (Sireci and Zenisky 2006). Furthermore, complex scoring methods make it possible to include both students' responses and decision making processes (e.g. Hao et al. 2016). In addition, recent new insights in adaptive algorithms could help to develop personalized learning and assessment systems. Thus meeting the increased need for personalization in both learning and assessment. All these innovations are promising in a sense that they can improve the quality of assessments significantly. Caution is required, however,

S. Wools (✉)
Cito, Arnhem, The Netherlands
e-mail: saskia.wools@cito.nl

M. Molenaar
Open Assessment Technologies, Luxemburg, Luxemburg

D. Hopster-den Otter
Universiteit Twente, Enschede, The Netherlands

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_1

since these innovations can also negatively impact important values of testing such as comparability and transparency.

Innovations in computer-based assessment can be described in a context of validity. Validity is one of the most important criteria for the evaluation of assessments (AERA, APA and NCME 2014) and is often defined as the extent to which test scores are suitable for their intended interpretation and use (Kane 2006). This chapter aims to address general aspects of computer-based assessment that can guide future validation efforts of individual computer-based assessment for a particular purpose and interpretation.

Validation efforts can be structured according to the argument-based approach to validation (Kane 2006, 2009, 2013), which is a general approach that can be used as a framework to structure validity evidence. The argument-based approach to validation aims to guide validation efforts by proposing a procedure that consists of two stages: a developmental stage and an appraisal stage. In the developmental stage, an interpretation and use argument (IUA) is constructed by specifying all inferences and assumptions underlying the interpretation and use of a test score. In the appraisal stage, a critical evaluation of these inferences and assumptions is given within a validity argument.

The IUA is structured according to several, predefined inferences (Wools et al. 2010): scoring, generalization, extrapolation, and decision. Every inference holds its own assumptions and underlying claims to argue valid use of test scores. When computer-based assessments are used, several additional claims are made—and subsequently, must be validated. At the same time, innovations in computer-based assessments can provide us with additional data or evidence that can support the validity of these assessments.

This chapter aims to describe and discuss several innovations in computer-based assessment from a validity perspective. The central question is: what are the threats to, and opportunities for, innovative computer-based assessments in the context of validity? In the first section, we describe the trends and innovations regarding computer-based assessments. These can be categorized into three categories: innovations in items or tasks; innovations in test construction, assembly and delivery; and innovations that accommodate students personal needs and preferences. The second section introduces the concept of validity and validation more thoroughly and describes the inferences underlying validity arguments. The two sections come together in the third, where the impact of technological innovation is discussed to establish the effect on the inferences from the argument-based approach to validation. In this section, we argue that these technological advancements are both improving as well as threatening the validity of assessments. And we propose some research questions that should be posed during the validation of innovative computer-based assessment.

1.2 Innovations in Technology-Enhanced Assessments

The use of technology in education is increasing significantly, access to the internet is ubiquitous, schools adopt new digital tools and students bring their own devices to the classroom. These technological advancements are not only limited to learning materials, also assessment can benefit. For example, when audio and video are used to create a rich and authentic assessment context that is appealing to modern-day students (Schoech 2001). Or, when process data and response times are gathered to improve insights in the behaviour on individual items (Molenaar 2015). These techniques can be used to further improve computer-based assessment of skills and competences. New technology can also help measure skills that were hard to measure by traditional CBA's. For example, previously, speaking ability could be measured through recording of speech, but scoring was done manually. Nowadays, cloud computing allows for AI-based automated scoring that was not possible before (Zupanc and Bosnic 2015). Technology can also be used to measure "new" competences like 21st century skills (Mayrath et al. 2012). As an example, assessing "collaborative problem solving" requires new types of items that include inter-agent interaction (OECD 2017). Digital technology makes it possible to create these types of items that go beyond the limits of what can be tested on paper with traditional multiple choice and constructed response interactions.

New (types of) devices and peripherals are being introduced at a rapid pace. The first Apple iPhone was introduced in 2007 and revolutionized mobile, personal computing and touch-based input. In 2009, Fitbit introduced the concept of wearable computing or "wearables", which has since evolved and branched out into head mounted displays (Google Glass 2013, Oculus Rift 2016) and smart watches (Google Watch 2014, Apple iWatch 2015). The iPad popularized the tablet in 2010 and received instant appeal from educators based on its friendly form factor and ease of use. In 2012, Microsoft's 2-in-1 Surface bridged the gap between tablets and laptops, appealing to audiences in higher education. And most recently smart speakers like Amazon Alexa (2014) and Google Home (2016) truly introduced us to the age of the assistant.

These devices have introduced new form factors, new types of input (interactions) and new output (sensor data). Natural touch/gesture based input has made technology more usable, allowing even infants to use it. Mobile phones have made audio (speech) and video recording accessible to all. And geographic, accelerometer and gyroscope data allow for more natural interactions with devices, localization and improved ease-of-use.

At the same time, the ubiquity of the internet allows for access to tools and information anywhere and at any time. Cloud computing has propelled machine learning. Providing us with endless possibilities, like on-the—fly video analysis to detect suspicious behaviour in airports or which groceries are put in shopping baskets (Johnston 2018). Voice assistants can send data to cloud-based algorithms to process natural language and follow-up on the requests of the user, including making reservations at a restaurant by a bot indistinguishable from a real-life person (Velazco

2018). Even in the area of creativity, AI has demonstrated being capable of creating artworks and composing music (Kaleagasi 2017).

This chapter discusses several practical implementations in the area of educational assessment today. Although there are many more technological innovations that impact assessment practices, examples are chosen within three distinct categories or levels:

1. Items and tasks: innovations in individual item and task design, ranging from simulations to multi-media enhanced items to hybrid tasks, (partially) performed in the real world
2. Test construction, assembly and delivery: innovations in automated item generation, adaptive testing and test delivery conditions by use of (online) proctoring
3. Personal needs and preferences: innovations to adapt to the personal needs and preferences of the individual student, ranging from accessibility tools to Bring Your Own Device (BYOD) to personalized feedback and recommendations in the context of learning.

1.2.1 Innovations in Items and Tasks

In educational measurement, technological innovations seem promising to support the assessment of "new" competences such as collaborative problem solving, creativity or critical thinking. These constructs are typically assessed through complex and authentic tasks. When these tasks are developed leveraging the possibilities of new technologies, new types of items emerge. These items have become increasingly more complex and are often referred to as Technology Enhanced Items (TEIs) (Measured Progress/ETS Collaborative 2012, p. 1):

> Technology-enhanced items (TEI) are computer-delivered items that include specialized interactions for collecting response data. These include interactions and responses beyond traditional selected-response or constructed-response.

By using these item types, it is possible to include sound and video and animations within the assessment. At the same time, performances are measured more direct and authentic. Finally, these items provide us with the opportunity to gather data beyond a correct or incorrect response. This additional data includes for example log-files, time stamps and chat histories. In general, challenges of TEIs are typically that they might favor digital natives, are harder to make accessible, are more expensive to develop in comparison with traditional items and that the resulting data are harder to analyze than with classical approaches.

As an example, we distinguish three types of TEI. The first TEI is a *simulation.* This is a digital situation where a student can roam a virtual environment and complete relevant tasks (Levy 2013). These simulations could be simple apps rendered within the context of a classic test to full immersive environments enriched by use of head-mounted VR headsets. Simulations are developed to simulate a specific situation that

invites students to respond to in a particular way: often these simulations are used to simulate an authentic situation.

The second type of TEI is one that can be used for video and audio recording (OAT 2018), using the student's device to record speech or capture video: *multi-media enhanced items*. These items are used to gather data that goes beyond constructed responses or multiple choice items. The speech and video fragments that are collected collected could be routed to either manual scorers or automated scoring algorithms (Shermis and Hammer 2012).

The last TEI is referred to as a *hybrid tasks.* These tasks are (in part) performed in the real world and can be directly (automatically) evaluated, allowing for greater interactivity. An example of a hybrid task is solving a table-top Tangram puzzle, which is recorded by a tablet-webcam and instant feedback is provided. (e.g., Osmo Play 2017).

1.2.2 Innovations in Test Construction, Assembly and Delivery

Regarding test construction activities, such as assembly and delivery, digital technology allows for automated item generation, adaptive testing and online (remote) proctoring for enhanced test security.

Automated item generation is the process of generating items based on predefined item models (Gierl 2013). These models can be very complex, taking into account the required knowledge and skills, but also personal preferences of students to make items more appealing, e.g. by posing questions in the context of topics personally relating to students like football or animals. This process can take place a priori to generate thousands of items to bootstrap an itembank, but also in real-time to adapt to personal preferences, apply digital watermarking for detecting origins of itembank-leaks (Foster 2017) and/or take into account real-time test-delivery data (responses, scores, process data).

Computer Adaptive Testing (CAT) has been around for a long time and is described as the process where test construction and test administration are computerized and individualized (Eggen 2007). There are many different types of CAT, each with their own objectives and technological implementations, but all with an adaptive engine that is used for real-time adaptive test assembly. From a technology standpoint, CAT can benefit from advancements in cloud computing, allowing real-time (re)calibration and even more complex computations and constraints to be evaluated. CAT-engines can also be designed to take into account prior data, demographics and personal needs and preferences. Prior data could be anything from previously selected items to the result of a previous test to an ability estimate by a teacher, in order to select a better first set of items to be administered, as that can increase efficiency significantly (van der Linden 1999). Demographics and personal needs and preferences are specific to the individual and provide instructions to the adaptive algorithm to balance content (e.g. exclude a certain topic) and take into account accessibility needs (e.g. exclude

items with images unsuitable for people suffering from color-blindness) or even device preferences (e.g. do not select items with drag & drop as this user is using a mouse).

On the test-delivery level, *online proctoring* allows for secure testing on any location, providing greater flexibility on where and when tests can be delivered. An adaptive test could be delivered in the comfort of a student's own home, while online proctors would proctor the test remotely by webcam surveillance. Leveraging cloud computing, real life proctors could be assisted by AI to process the video data detecting aberrant behavior (e.g. sudden movements or voices in the background). And real-time data forensics engines could be used to spot anomalies during the actual test taking, e.g. answering items correctly at a very high speed or suspicious answers patterns indicating possible collusion with other students.

1.2.3 Innovations Regarding Personal Needs and Preferences

Lastly, digital technology allows assessments to be adapted to personal needs and preferences. Personal needs and preferences are typically related to (legal) accessibility requirements, but can also be personal preferences of any kind, e.g. a preferred type of device (tablet, laptop) or screen size. The latter is closely related to the phenomenon of Bring Your Own Device (BYOD), where students bring their own devices into the classroom and want to perform their tasks using the device and configuration they are familiar with. Also, when personal needs and preferences are saved, it becomes possible to present students with personalized feedback.

Tools for Accessibility are products, devices, services, or environments for people with disabilities and are becoming increasingly important in the area of assessment, to ensure all students have equal opportunities when taking a test. The foundations and the extent of accommodations may vary but in many countries it is simply required by law (e.g., American Disability Act). Also the types of accommodations can vary, ranging from always-on accessibility features, to extended features based on personal profiles and to the use of specialized Assistive Technologies like screen readers or refreshable braille devices.

Another type of personalization is the use of the preferred *device* type or form factor trough *BYOD* (*bring your own device*). Typical web applications employ the principle of responsive design: an approach to web design that makes web pages render well on a variety of devices and window or screen sizes. Based on device capability and available screen estate, content is reformatted dynamically to provide the best possible user experience. Apart from screen estate, also available input types can play an important role, e.g. a mobile phone with an on-screen keyboard, a tablet with a type cover or a laptop with a physical keyboard can yield different results (Laughlin Davis et al. 2015). Some students may be very proficient with a certain input type, whereas others might struggle. To accommodate for this, it is important for students to either be able to use the (type of) device of their preference or are allowed sufficient time to practice and get acquainted with the compulsory/recommended device and mode.

Personalization transcends the process of assessment delivery; the type and mode of feedback can also be individualized, taking into account personal preferences, prior data and learning styles. This *personalized feedback* constitutes to personalized or adaptive learning, where an AI-based recommendation engine can take into account all factors and data to provide the appropriate type of the feedback and content, tailored to the exact needs of the individual: e.g. the next chapter to read, video to watch or exercise to complete.

1.3 Validity and Validation

Extensive research caused the concept of validity to change significantly over time (Lissitz 2009). Kane (2006) summarized this by citing three general principles of validation that emerged from the widely accepted model of construct validity (Cronbach and Meehl 1955). The first principle concerns the need to specify the proposed interpretation of test scores. The second principle refers to the need for conceptual and empirical evaluation of the proposed interpretation. The third principle states the need to challenge proposed and competing interpretations. All these principles are reflected in widely known theories on validity and approaches to validation. For example, in Messick's (1989, p. 13) definition of validity:

> …an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment [italics in original].

Messick's conceptualization of validity has resulted in a validation practice that aimed to present as much validity evidence as possible. From this practice, the validity of test scores has been supported by combining countless sources of validity evidence that are either content-related, criterion-related, or construct-related. To propose a more pragmatic practice, Kane suggested the argument-based approach to validation (2004, 2006). This approach guides validation efforts through selecting the most relevant sources of evidence and therefore lessens the burden on practitioners. According to Kane (2013, pp. 8–9):

> The argument-based approach was intended to avoid the need for a fully developed, formal theory required by the strong program of construct validity, and at the same time to avoid the open-endedness and ambiguity of the weak form of construct validity in which any data on any relationship involving the attribute being assessed can be considered grist for the mill (Bachman 2005; Cronbach 1988; Haertel 1999; Kane 1992).

The argument-based approach to validation consists of two arguments: an interpretation and use argument (IUA) and a validity argument. The IUA states which inferences and assumptions underlie the intended interpretation and use of test scores. Whereas the validity argument evaluates the evidence that is presented to support or reject the inferences from the IUA and draws a conclusion on the adequacy of the validated instrument for the intended interpretation and use.

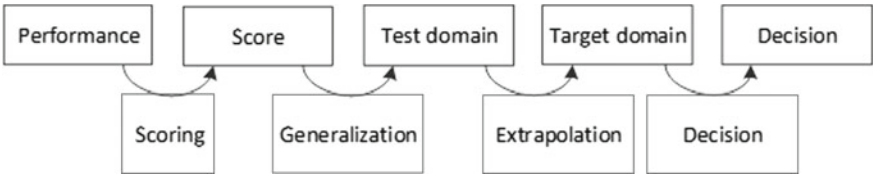

Fig. 1.1 Inferences within an IUA (Kane 2006)

When looked at in more detail, the IUA helps us to specify our reasoning from an observed performance in an assessment situation towards a decision and the use of this decision for a particular purpose (e.g., selection, classification, a didactical intervention), this is represented in Fig. 1.1. The first inference (scoring) describes how a students' performance on tasks is translated into an observed score. These scores are usually interpreted as a generalizable instance of a test domain score (generalization). A test domain represents all possible tasks that could be presented to students within the chosen operationalization of the construct. The test domain scores are subsequently extrapolated (extrapolation) to scores on a broader domain. This domain can be a theoretical competence domain, which entails an operationalization of the competence or construct that is being measured. It can also entail the practice domain, that is, a real-life situation that students can encounter in their future (professional) lives. Whether the test domain scores are extrapolated into a theoretical competence domain or a practice domain depends on the particular testing situation, in general one could say that we extrapolate to a target domain. Either way, building on this extrapolation, the final inference (decision) can lead to a decision on the students' level on the competence of interest.

After developing the IUA, analytical and empirical evidence are gathered to enable an evaluation of the claims stated in the IUA. The analytical evidence could entail, for example, conceptual analyses and judgments of the content of the test domain and competence domain. Most of the analytical evidence could already have been generated during development of the assessment. The empirical evidence consists of data relating to, for example, the reliability of an assessment, the structure of the construct, or relations with other measures of the construct of interest. This kind of evidence is gathered in so-called validation studies, which are designed to answer specific research questions derived from the need for specific empirical evidence. The evidence is used for the validity argument, that includes a critical evaluation of the claims in the IUA. Note that this consists of both appraising currently defined inferences and assumptions and rejecting competing interpretations.

One might think that the argument-based approach encourages to gather all possible analytical and empirical evidence. However, according to Kane (2009, p. 49):

> …some statements in the literature can be interpreted as saying that adequate validation requires that every possible kind of validity evidence be developed for validation to be complete.…

This shotgun approach is clearly unwieldy, and in its extreme form, it makes validation impossible.

Therefore, within the argument-based approach, it is argued that inferences that seem weak or that are of great interest to the intended interpretation and use of tests require more evidence than others. Although evidence is needed for every inference, the weight placed on different inferences depends on the assessment that is being validated.

The argument-based approach to validation is applied to several assessments and, when necessary, adapted to fit different perspectives or uses of tests. The most prominent shift in the approach was proposed by Kane (2013) when he argued that the use of assessment results should play a more prominent role in the approach. This resulted in a change in terminology: the theory moved from using an interpretive argument into using an interpretation and use argument. Others, also published specific applications of the argument-based approach or proposed extensions to parts of the theory: for language assessments (Llossa 2007), for assessment programs (Wools et al. 2016), for classroom assessment (Kane and Wools 2019) and for formative assessment (Hopster-den Otter et al., submitted).

The current chapter applies the argument-based approach to innovative computer-based assessments. Coming back to the subject of this chapter, when developing or validating innovative computer-based assessments, one might need to rethink which inferences seem weak or are of great interest. Also, when validation practice is not moving along with digital innovations, we might target our validation efforts at the wrong inferences. Therefore, in the remainder of the chapter we will give an overview of the impact of innovations in computer-based assessments and where they might impact our claims and assumptions underlying the inferences in the IUA.

1.4 Validity of Innovative Technology-Enhanced Assessments

The argument-based approach to validation starts with specifying inferences, assumptions and claims that are made in the assessment process. Since innovations in computer-based assessments have impact on all aspects of the assessment process, the impact on validity is large. In this section we will discuss the inferences distinguished in an IUA. From there, we discuss specific claims, assumptions and threats underlying the inferences when technological enhanced innovations are used in assessment. This provides an overview of possibilities and threats within a validity argument that should play a central role when gathering and evaluating validity evidence.

1.4.1 Inferences Within the IUA

As previously mentioned, an IUA consists of a set of inferences and accompanying claims and assumptions. It depends on the particular assessment and the intended

interpretation and use of the assessment scores what claims and assumptions are relevant within an IUA. We exemplify the inferences in general with claims and inferences that are commonly used (Wools et al. 2016).

Scoring inference

When students perform an assessment task, such as answering items or solving a complex problem, data are collected to transform the students' behavior into an interpretable unit. Usually this is a score that indicates whether the answer was correct, or if possible, partially correct. This inference implies that it is possible to make statements about a task being performed correctly or not. Another assumption is that the score is a true translation of students' ability to perform on the task. The final assumption underlying the scoring inference is that students are able to show their skills or competences without barriers. In practice, this means that students know what is expected of them, that the tools work intuitively, and that they are able to perform the task without technological difficulties.

Generalization inference

Generalizing a score from an assessment to a test domain means that the responses on that particular assessment can be interpreted as representative for all possible tasks or test forms that could have been presented. It also means that the performance must be more or less the same when a student takes the test twice with different items. This implies that the tasks in one assessment must be representative for the full test domain and that this is comparable for different versions or instances of the assessment.

Extrapolation inference

When we extrapolate a score on the test domain to a target domain we assume that the tasks within the test domain are derived from this target domain. This inference relies very heavily on one claim: the task is as authentic as possible. When assessment tasks are very authentic, the extrapolation of what we observed is not far from what we would like to make decisions about.

Decision inference

The main question for this inference is: are we able to make a decision about students that is in concurrence with the intended interpretation and use of the assessment? This implies that we have meaningful cut scores or norms that can be applied to students performances. Furthermore, it is implied that the results are meaningful to students and that they can be used for the intended purpose.

1.4.1.1 Innovations in Items and Tasks

When TEI's are used, it is possible that defining the correct response becomes more complex. Different processes, responses or answers could all be considered effective behavior and therefore assumed to be 'correct'. Furthermore, translating behavior

into a single score does not always reflect the effort that has been put into the tasks. Therefore, to quantify behavior on these new types of tasks, data that describe the followed process (log-files) are often collected and used for analysis and reporting. This means the use of complex algorithms to score behavior or the use of, for example, automated essay scoring to evaluate the quality of an answer. The risk with these algorithms is that scoring becomes less transparent and hard to verify, especially when machine learning is used and so called black-boxes are created. This threatens the *scoring inference* in a way that it becomes harder to evaluate whether a score is given correct.

The *generalization inference* assumes that tasks are selected to cover all relevant aspects of a construct. A risk for innovative technology enhanced assessments is construct underrepresentation. Construct underrepresentation occurs when only a small aspect of a construct is assessed. For example, we assess the ability to converse about the weather in another language while the intend was to make a decision about someone's full ability of conversing in another language. In technology enhanced assessment, TEIs are often used. However, developing these tasks is a time consuming and costly effort that leads to a limited set of contexts or tasks. Moreover, time constraints during the administration of the test, or other practical limitations in the administration, often prevent us from presenting a large number of tasks, contexts and skills. When limited items are presented, this will threaten the generalizability of the obtained scores to the full test domain.

At the same time, these TEI's provide us with more opportunities to build rich and authentic tasks. Simulations include open digital environments where a student can virtually roam and complete relevant tasks. Items that include multi-media provide the possibility to grasp performance on video or record speech. And finally, hybrid tasks invite students to perform offline (for example solve a puzzle) and provides them with online feedback. This last examples makes sure that the computer is not 'in between' the student and his or her performance anymore. All in all, all these tasks are developed to provide the candidate with an authentic experience. This way the behavior that is called for in the assessment situation is as identical as possible as the behavior that requested in the competence domain. Therefore, through these authentic items the *extrapolation inference* is strengthened.

The *decision inference* includes assumptions about cut scores and norms. Cut scores and norms are usually the result of statistical analysis, equating techniques or standard setting procedures. Unfortunately, the commonly used methods are not always suitable for the rich data that are produced through TEI's. This means that even when these tasks are scored in a transparent, reliable and comparable way, it might still be a problem to decide 'what behavior do we consider good enough to pass the assessment?'.

1.4.1.2 Innovations in Test Construction, Assembly and Delivery

Within the *scoring inference*, it is assumed that a score assigned to a student's performance is a translation of a student's ability. More specifically, that the score is only

influenced by the performance of a student on the task at hand and not, for example, by other students who could help. Therefore, this assumption does not hold when students discuss their answer with others. Fortunately, cheating becomes harder to do with new security measures like (online) proctoring. Furthermore, adaptive algorithms and live test assembly allow for individual test forms, making copying of answers more difficult.

The *generalization inference* is concerned with reliability and comparability between test forms in terms of content representation. In terms of comparability between test versions, adaptive engines can be used to make sure different test versions are comparable in terms of content. These engines use sophisticated rules to sample items within certain content restrictions. To be able to do this, the item pool must be large enough. If this would be the case, then content comparability can be ensured over different versions and therefore, these engines strengthen our generalization inference.

As mentioned previously, authenticity is an important aspect of the *extrapolation inference*. One of the threats to authenticity is the availability of tasks that speak to a candidates personal interests. For example, an authentic situation for a student to read texts, is often to read a text that holds information that a student is interested in. Advances in automated item generation support test developers in constructing items that can speak to different personal interests of students. Therefore, it is possible that AIG can positively influence authenticity and therefore extrapolation.

The decision inference is concerned with cut scores, norms and score reports. The innovations mentioned regarding test construction, assembly and delivery are not related to these aspects and this inference.

1.4.1.3 Innovations Regarding Personal Needs and Preferences

An assumption underlying the *scoring inference* is that students are able to answer items or perform tasks without barriers. For example, when students need to work with formulas, they should not be limited to demonstrate their skills because of the complexity of the formula-editor that is used in an assessment context. This can also occur when a device that is used in the testing condition is different from the one a student is used to. As an example, someone who is used to an Android phone usually has problems in using an iPhone and the other way around. In an assessment situation these unnecessary difficulties cause for construct irrelevant variance since the score does not reflect the true ability of the student anymore, but is impacted by the ability to cope with other technical devices. One of the solutions in current assessments is a Bring Your Own Device (BYOD) policy where students can use devices and tools that they worked with during the learning phase. For students with special needs, this means that they can also use their own tools for accessibility, such as screen reader software or a refreshable braille device. We acknowledge that this strengthens the claim that underlies the scoring inference, but at the same time, it raises questions about comparability and might weaken other inferences.

The generalization inference is an example of an inference that might be weakened through a BYOD policy or allowing tools for accessibility. This is, when students bring their own devices and use their personal software and peripherals to ensure accessibility, the claim regarding comparability is challenged. Especially when items or tasks are not suited for these different modes (touch screen devices vs. mouse-controlled devices) or when item presentation varies over different devices. This causes items to differ in terms of necessary cognitive load and therefore in their difficulty for students.

To strengthen the *extrapolation inference*, we would like the assessment situation to be as authentic as possible. For example, when we want to say something about someone's ability to sing a song at home—the most authentic way is to let them sing a song. However, there are still aspects that would prevent us from being able to make that claim. What if the person is really nervous when there is an audience or when it is necessary to use a microphone? Therefore, even if a task is authentic, there is still an inference to be made about the possibility to extrapolate the observed behavior into potential behavior on the target domain. As mentioned before, it becomes more common for students to be able to use their own device and tools that they are used to work with during class. This bridges some of the extrapolation gaps between the assessment context and learning situation and therefore could positively impact the extrapolation inference.

When an assessment is over, a decision about students is made. Within the decision inference it is assumed that is possible to give meaning to the test results. This is done through norms, cut scores and score reports. These score reports usually consist of a visualization of the assessment score interpretation. However, it can also include feedback that is aimed to guide students' further learning. An advantage of technological advancement is the possibility to provide students with personalized feedback. The feedback that should be presented is not only selected based on item answers and ability estimates, but can also be selected based on analysis of learning patterns and learning preferences. When this is done in a formative assessment context or a context of classroom assessment, assessment results are translated into meaningful actions right away, strengthening the claim that the results can be used for the intended purpose. This is not only the case for formative assessment, also for summative assessment or classification purposes, the possibility to combine different data sources to decide on the next best step strengthens the inference.

1.4.2 Validity Argument of Technology-Enhanced Assessments

A validity argument consists of an integral evaluation of all sources of evidence and a critical appraisal of the claims and inferences. This is necessary because a design choice or a particular source of evidence can support an inference and at the same time threaten another. For example, the use of simulation-based assessment might be

Table 1.1 Opportunities (+) and threats (−) for validity

	Scoring	Generalization	Extrapolation	Decision
Items and tasks				
Simulations	−	−	+	−
Multi-media enhanced tasks	−	−	+	−
Hybrid tasks	−	−	+	−
Test construction, assembly and delivery				
Automated item generation			+	
Adaptive engines	+	+		
(online) Proctoring	+			
Personal needs and preferences				
Tools for accessibility	+	−	+	
Bring your own device	+	−	+	
Personalized feedback				+

evaluated positive in light of the extrapolation inference, but gives reason for concern for the scoring and generalization inference. Table 1.1 shows this in more detail. The technological innovations discussed in Sect. 1.2 of this chapter are listed as well as the four inferences. For every inference it is noted whether an innovation has the potential to be an opportunity (+) or a threat (−). Some technological innovations were not discussed in relation to the inference or are not applicable and are left empty.

The validity argument aims to present a balanced case to come to a conclusion about the overall validity of the test scores. When the results from Table 1.1 are taken into account, it stands out that the evidence most needed for innovative assessment is to strengthen the scoring, generalization and decision inference. Questions that should be addressed are, for example, can innovative items be scored in a transparent way that includes all relevant aspects of the task? Is the sample of behavior representative enough to justify claims that go beyond the assessment situation? Is comparability between test versions and test conditions plausible? And is it possible to make decisions about students performances that are both explainable and meaningful?

1.5 Concluding Remarks

In this chapter, we discussed the impact of several technological innovations from a validity perspective. Validity and validation are defined from a perspective of the argument-based approach and according to this approach several inferences are distinguished. For every inference, claims were specified and related to a limited set of technological trends from educational assessment practice. Some of these practices are strengthening the validity claims, others are weakening them.

In general, one could say that the scoring and generalization inference are of major concern when using these innovative techniques. Also, the use of innovative assess-

ment tasks, such as simulations, multi-media enhanced tasks or hybrid assessment tasks is quite ambiguous from a validity point of view: it strengthens the extrapolation inference, but weakens the scoring, generalization and decision inference. It is important to note that this could be solved relatively easy by providing evidence that rejects the assumptions of incomparability between tasks or shows how these innovative tasks can be scored. An advantage of the technological advancement in this context, is that the new data that these tasks provide us with, and the new data analysis techniques that are available, can be of help in creating this evidence and to study these claims more extensively.

Furthermore, we stress that this is a general reflection of an IUA. For every assessment it is necessary to build a custom IUA specifying the inferences, claims and assumptions relevant to that assessment and its intended interpretation and use of the test scores. Moreover, every assessment probably holds its' own unique combination of innovative features that might combine differently than the ones presented here. Building an interpretive argument for a specific computer-based assessment is helpful in deciding what the weakest inferences are and where to target validation efforts on.

One thing that stands out, is that many innovations are practice-based. Test developers, assessment organizations, start-ups and even technological companies develop techniques to improve assessment. However, little are evaluated and reported on in a systematic way, let alone published in scientific journals. This is an important aspect of validity, the use of evidence-based techniques that are tested and proven lessens the burden to gather additional data and strengthens our claims. Also, when innovations or new techniques from other fields are used in assessment, it is necessary to study and publish on the impact of these advancements.

We acknowledge that there are many more advancements that were not discussed. Some of these will strengthen and some will weaken validity. Also, some of these innovations are very expensive to build or use, others are easy to implement. We conclude that a lot of technological advancements hold potential to improve assessments considerably, we should, however, not forget that a lot of assessment-challenges can be addressed by traditional types of assessments and items as well. Only when it is necessary, these innovations might be able to truly add value. And when it seems that these innovations bring new problems that seem unsolvable, keep in mind: traditional assessments have problems of their own, those are simply the problems we are familiar with by now.

References

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Bachman, L. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*, 1–34.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Eggen, T. J. H. M. (2007). Choices in CAT models in the context of educational testing. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*.

Foster, A. (2017, June 20). *What national security can teach us about protecting our exams*. Retrieved from https://www.caveon.com/2017/06/21/what-national-security-can-teach-us-about-protecting-our-exams/.

Gierl, M. (2013). *Advances in automatic item generation with demonstration* [PowerPoint slides]. Retrieved from https://www.taotesting.com/wp-content/uploads/2014/09/TAO-Days-AIG-October-2013.pdf.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5–9.

Hao, J., Smith., L., Mislevy, R., Von Davier, A., & Bauer, M. (2016). *Taming log files from game/simulation-based assessments: Data models and data analysis tools*. ETS Research Report Series, 1–17. https://doi.org/10.1002/ets2.12096.

Hopster-den Otter, D., Wools, S., Eggen, T. J. H. M., & Veldkamp, B. P. (2018). *A general framework for the validation of embedded formative assessments*. Manuscript submitted for publication.

Johnston, C. (2018, January 22). *Amazon opens a supermarket with no checkouts.* Retrieved from https://www.bbc.com/news/business-42769096.

Kaleagasi, B. (2017, March 9). *A new AI can write music as well as a human composer: The future of art hangs in the balance.* Retrieved from: https://futurism.com/a-new-ai-can-write-music-as-well-as-a-human-composer.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535.

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

Kane, M. T. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 39–64). Charlotte, NC: Information Age Pub.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000.

Kane, M. T., & Wools, S. (2019). Perspectives on the validity of classroom assessments. In S. Brookhart & J. McMillan (Eds.), *Classroom assessment and Educational measurement.* Abingdon, Oxon: Routledge.

Laughlin Davis, L., Kon, X., & McBride, Y. (2015). *Device comparability of tablets and computers for assessment purposes.* Paper presented at National Council on Measurement in Education, Chicago.

Levy, R. (2013). Psychometric and evidentiary advances, opportunities, and challenges for simulation-based assessment. *Educational Assessment, 18,* 182–207. https://doi.org/10.1080/10627197.2013.814517.

Lissitz, R. W. (2009). *The concept of validity: Revisions, new directions, and applications*. Charlotte, NC: Information Age Publishing INC.

Llosa, L. (2007). Validating a standards-based assessment of English proficiency: A multitrait-multimethod approach. *Language Testing, 24,* 489–515. https://doi.org/10.1177/0265532207080770.

Mayrath, M. C., Clarke-Midura, J., Robinson, D. H., & Schraw, G. (Eds.). (2012). *Technology-based assessment of 21st century skills: Theoretical and practical implications from modern research*. Charlotte, NC: Information Age Publishing.

Measured Progress/ETS Collaborative. (2012). *Smarter balanced assessment consortium: Technology enhanced items*. Retrieved from https://www.measuredprogress.org/wp-content/uploads/2015/08/SBAC-Technology-Enhanced-Items-Guidelines.pdf.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: Macmillan.

Molenaar, D. (2015). The value of response times in item response modeling. *Measurement: Interdisciplinary Research and Perspectives, 13*(3–4), 177–181. https://doi.org/10.1080/15366367.2015.1105073.

OAT. (2018). *Userguide TAO—Portable custom interactions*. Retrieved from: https://userguide.taotesting.com/3.2/interactions/portable-custom-interactions.html.

OECD. (2017). What is collaborative problem solving? In *PISA 2015 results* (Vol. V)*: Collaborative problem solving*. Paris: OECD Publishing. https://doi.org/10.1787/9789264285521-7-en.

Osmo Play. (2017). Retrieved from: https://www.playosmo.com/en/tangram/.

Schoech, D. (2001). Using video clips as test questions: The development and use of a multimedia exam. *Journal of Technology in Human Services, 18*(3–4), 117–131. https://doi.org/10.1300/J017v18n03_08.

Shermis, M., & Hammer, B. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. In *Annual National Council on Measurement in Education Meeting* (pp. 1–54).

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Smarter Balanced Assessment Consortium. (2012). *Technology-enhanced items guidelines*. Developed by Measured Progress/ETS Collaborative. Retrieved from: https://www.measuredprogress.org/wp-content/uploads/2015/08/SBAC-Technology-Enhanced-Items-Guidelines.pdf.

Van Der Linden, W. J. (1999). Empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement, 23*(1), 21–29. https://doi.org/10.1177/01466219922031149.

Velazco, C. (2018, June 6). *Google's reservation-making AI will be making calls soon.* Retrieved from: https://www.engadget.com/2018/06/27/google-duplex-assistant-public-testing/?guccounter=1.

Wools, S., Eggen, T. J. H. M., & Béguin, A. A. (2016). Constructing validity arguments for test combinations. *Studies in Educational Evaluation, 48,* 10–16. https://doi.org/10.1016/j.studeduc.2015.11.001.

Wools, S., Eggen, T. J. H. M., & Sanders, P. (2010). Evaluation of validity and validation by means of the argument-based approach. *CADMO, 8,* 63–82.

Zupanc, K., & Bosnic, Z. (2015). Advances in the field of automated essay evaluation. *Informatica* (Slovenia), 39.