Getting the First and Second Decimals Right: Psychometrics of Stealth Assessment

Seyedahmad Rahimi[1]*, Russell G. Almond*[2]*, Valerie J. Shute*[2]*, & Chen Sun*[2]

[1]*University of Florida,* [2]*Florida State University*

Abstract

Stealth assessment, like all assessments, must have three essential psychometric properties: validity, reliability, and fairness. Evidence-centered assessment design (ECD) provides a psychometrically sound framework for designing assessments based on a validity argument. This chapter describes how using ECD in the design of a stealth assessment helps designers meet the psychometric goals. It also discusses how to evaluate a stealth assessment's validity, reliability, and fairness after it is designed and implemented.

Keywords: Stealth assessment, evidence-centered design, psychometrics, validity, reliability, fairness

Introduction

The field of psychometrics in education is mainly concerned with quantitative and qualitative methods, techniques, and guidelines leading to designing and developing high-quality measurements and assessments. According to Messick (1994), "…validity, reliability, comparability, and fairness need to be uniformly addressed for all assessments because they are not just measurement principles, they are social values that have meaning and force outside of measurement wherever evaluative judgments and decisions are made" (p. 13). In this chapter, we focus on reliability, validity, and fairness in stealth assessment (Shute, 2011). First, we review what these three psychometric parameters mean, and later in this chapter we discuss common methods for evaluating these parameters.

Reliability refers to the consistency of an assessment. For example, a highly reliable bathroom scale shows a person's weight as about 140 lb. in the morning, afternoon, and evening. Conversely, a scale with low reliability shows that same person's weight as 140 lbs in the morning, 80 lbs. in the afternoon, and 190 lbs. in the evening. Reliability is an inherent property of a measurement. The question is not whether a measure is reliable, but whether it is sufficiently reliable for a given purpose (American Educational Research Association et al., 2014). The extent to which an assessment is reliable (consistent) can be evaluated using various techniques (e.g., correlations between two parallel test forms). We will discuss some of those techniques later in this chapter.

Validity refers to the extent to which an assessment is assessing what it claims to assess (Messick, 1994; Shute, 2009). Similarly, the *Standards for Educational Psychological Testing* (American Educational Research Association et al., 2014) indicates that, "validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests." (p. 11). One might say, "I am assessing creativity," but are they? If they are, how accurately are they assessing creativity (or any other competency), and how accurate are they interpreting the results of their assessment (Shute, 2009)? An alternative word for validity could be accuracy. Kane (2006) approaches validation as constructing a formal argument, accumulating evidence for why the scores support the proposed interpretation. There are several types of validity argument (e.g., content, construct, and criterion validity) and a complete validity argument will use multiple types. Note that reliability is a prerequisite for validity; an assessment cannot consistently measure the target construct (high validity) if it is not consistent (low reliability).

Fairness refers to the extent to which an assessment is equitable and unbiased for various subgroups (DiCerbo et al., 2016; Dorans & Cook, 2016; Mislevy et al., 2013). To say what is fair, one can start by saying what is not fair (Dorans & Cook, 2016). For example, from the assessment-design perspective, an assessment is not fair if it includes items that include culturally sensitive concepts seen as appropriate for some people and inappropriate for others. From the assessment-administration perspective, if an assessment requires certain equipment that some people have and others don't (e.g., the need to have a computer and Internet access for an assessment in a remote village in Africa), that assessment is not fair. Fairness can also be seen as a factor affecting the validity of an assessment. For example, suppose some construct-irrelevant variance exists in the assessment estimates/scores (e.g., the ability to work with a computer mouse affecting learners' score on a math assessment). In that case, both fairness and validity of the assessment are questionable (Dorans & Cook, 2016). Or suppose a literature exam was going to have questions on a Shakespeare play, and teachers from one school learned which play and instructed their students to read it before the test, but teachers from other schools did not. Then which school the student went to would relate to performance in a way that was independent of student learning.

To assess fairness, it is not sufficient to compare the average performance of different demographic groups on the assessment (Holland & Wainer, 1993). Statistically, any two groups almost never have exactly the same performance, and in the case where there has been a history of discrimination against one group, lower scores could be due to lack of access to resources. Thus, average score differences would be a sign that additional remedies are needed to correct resource availability problems. A better test for fairness would be to look at differential performance by individuals at about the same ability level but from different groups.

There are best practices that assessment designers can follow to make sure their assessment is valid, reliable, and fair. The purpose of this chapter is to explore these best practices in the context of stealth assessment, drawing examples from our work on a game called *Physics Playground* (*PP*; Shute et al., 2019).

Stealth Assessment

As discussed in other chapters of this book, stealth assessment is an assessment technique empowered by the evidence-centered design framework of assessment (ECD; Almond et al., 2015). Stealth assessment is usually embedded within a digital game but can be used in other types of interactive environments, so that the learner focuses on engaging with the environment and not on being assessed. The learner's interactions with the game environment are captured in log files, and the stealth assessment machinery processes and accumulates their activities into estimates ability level per targeted skill. These estimates are continuously updated as the learner progresses through the game, allowing the game to track current knowledge and skills, then adapt the environment accordingly. For instance, the game can adjust the difficulty of challenges, the sequencing of activities, and/or provide appropriate supports (e.g., cognitive or affective) based on each learner's current ability level. Although stealth assessment is intended to be an assessment *for* learning (i.e., formative assessment), it still can provide reliable and valid summative estimates of the competencies (i.e., assessment *of* learning) (e.g., Shute, Hansen & Almond, 2008).

Like traditional assessments, stealth assessment must be valid, reliable, and fair; using ECD as the design framework provides a mechanism for documenting the evidentiary argument underlying the assessment. We start by discussing ECD core models.

ECD Core Models

The evidence-centered design framework of assessment (ECD; Almond et al., 2015) facilitates the design of high-quality assessments. ECD includes four main models (see Figure 12.1). The Competency Model (CM) defines and displays the construct that we want to assess, which is usually latent and unobservable. The Evidence Model (EM) delineates the evidence that's needed to make claims, in terms of all targeted indicators, behaviors, or interactions done by the learners that can be observed and we call them *observables*—and asserts (a) how those indicators will be scored (evidence rules), and (b) how the scored evidence will be statistically linked and accumulated to inform the competency model. The Task Model (TM) describes the particular features of a situation or task which permit one to elicit the necessary evidence to assess the competency of interest. Task model variables are often related to the stimulus material provided with the task, for example, the length and complexity of the instructions. These variables are also associated with how the responses are collected, for example, whether the student selects an option from a list, or enters a free text response. Task description (which involves task model variables) include two main categories: (1) presentation materials (i.e., what will be shown to the learner, like a prompt or certain media), and (2) work products (i.e., learners' actions or responses that will be recorded in the log files). Task authors (game designers) manipulate these variables when building game levels. The ensemble of tasks must span both the breadth and depth of the target competencies. The Assembly Model (AM) describes the final collection of tasks and how they are sequenced. The AM ensures that sufficient evidence is collected for a reliable and valid assessment.

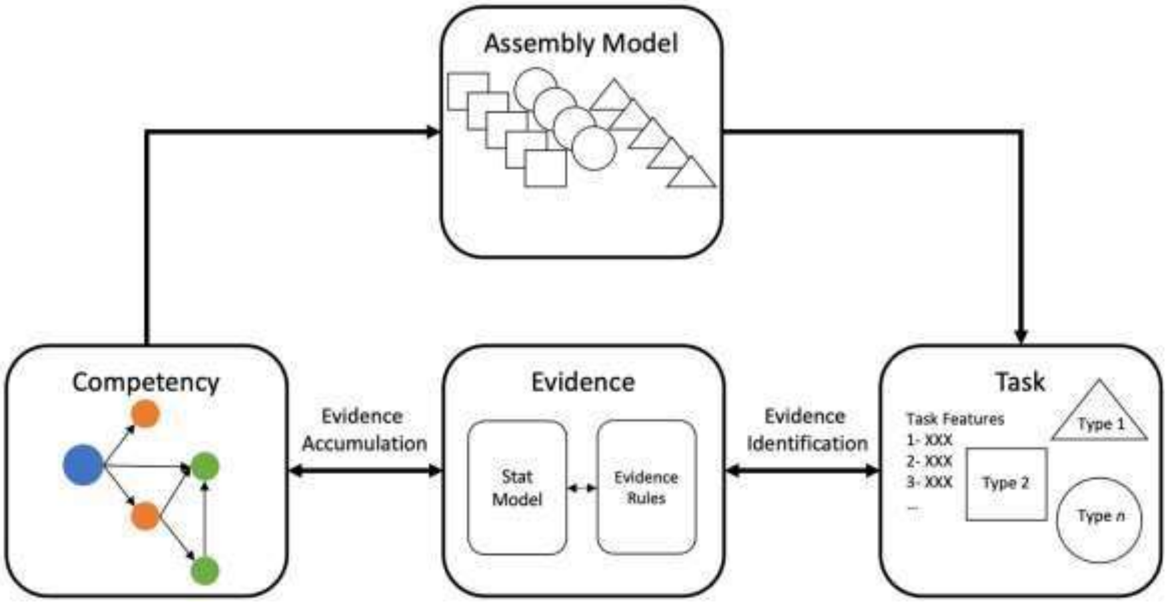

Figure 12.1: Four main models of an ECD (adapted from Mislevy et al., 2003)

The information in the ECD models becomes specification for the four processes used to implement an assessment (Almond et al., 2002; Almond, 2020). A digital game that uses stealth assessment must have capabilities for these four processes: (1) presentation and evidence capture (EC)—the game interface which captures (logs) learner interactions with the game, (2) evidence identification (EI)—the task-level scoring system which identifies the key observable outcomes (often from learner-generated log data; referred to as *observable[s]* in this chapter), (3) evidence accumulation (EA)—the part of the assessment system which combines observables from the EI process (task-level scores) and uses them to estimate competency levels (often using a statistical model such as Bayesian Networks, BN), and (4) activity selection (AS)—the algorithm responsible for selecting and sequencing the game levels as well as determining when enough evidence has been gathered to stop. Shute, Lu, and Rahimi (2021) listed nine steps to design and develop a stealth assessment, to which we have added a tenth (see Table 1).

Table 1
Stealth Assessment's Steps (adapted from Shute et al., 2021).

Step	**Qual / Quant**	**Description**	**ECD Model**
1	Qual	Develop competency model of targeted knowledge, skills, or other attributes based on full literature and expert reviews.	CM
2	Qual	Determine which game (or learning environment) the stealth assessment will be embedded into.	
3	Qual	Delineate a full list of relevant gameplay actions/indicators that serve as evidence to inform the CM and its facets.	EM
4	Qual	Create new tasks in the game, if necessary.	TM
5	Qual	Create Q-matrix to link actions/indicators to relevant facets of target competencies.	EM and AM
6	Qual	(Not in Shute et al., 2021) Decide on the collection of activities (both assessment tasks and learning activities), rules for navigating between them, and stopping rules for when there is enough information.	AM
7	Quant	Determine how to score indicators using classification into discrete categories (e.g., yes/no, very good/good/ok/poor relative to quality of the actions). This becomes the "scoring rules" part of the evidence model.	EM
8	Quant	Establish statistical relationships between each indicator and associated levels of the CM variables.	EM
9	Qual & Quant	Pilot test scoring model (e.g., BN) and modify parameters.	EM
10	Qual & Quant	Validate the stealth assessment with external measures.	
11	Quant	Use the current learners' competency estimates to provide adaptive game challenges or adaptive learning support.	

Note. CM = Competency Model, EM = Evidence Model, TM = Task Model, AM = Assembly Model, Quant = Quantitative, Qual = Qualitative.

Although sometimes omitted from summaries of ECD, the assembly model critically determines the reliability of an assessment. That is, an assessment is a collection of tasks for the learner to perform, and in a game-based assessment these are often levels or challenges in the game. The reliability of the game is determined both by the number and evidentiary strength of the challenges (if we have very few observations for a competency, we can't make a claim that our assessment is consistent). The best way to improve the reliability of an assessment is to manipulate the number and types of challenges offered. The validity of the assessment is established by producing a collection of challenges that span both the depth and breadth of the competencies being measured. If an assessment does not cover all the competencies (and their sub-facets), it will not be accurate. We discuss considerations related to fairness later in this chapter.

Note that ECD models can be specified at two layers of detail. The first is the more qualitative layer (called *domain modeling* in Mislevy, et al., 2003) which entails the basic evidentiary argument being laid out. In particular, the targeted competencies are given measurable definitions and linked to specific tasks that can provide evidence about them. The designers must then create a collection of challenges (game levels) that will provide evidence about the target competencies adequate to the purposes of the assessment (formative or summative). In the second, more quantitative layer (called the *conceptual assessment framework*), the actual statistical models and code used to implement the assessment must be specified. This includes both the population model for competencies in the target population and the statistical portions of the evidence models linking observations to competencies. Eventually, these statistical models can be refined or calibrated using data from pilot testing to improve the reliability and validity.

As reliability and validity are typically numeric values, the first domain modeling step determines the first digit of accuracy. Again, it is nearly impossible to get a reliable measurement with too few tasks, or a valid measurement with a collection of tasks that do not span the depth and breadth of the competencies. The second more quantitative step involves choosing a statistical model for scoring, which comprises the second digit of accuracy. Elaborate mathematical modeling cannot save an assessment that does not contain the right collection of tasks. However, the quantitative design can build on a good qualitative design foundation to improve the accuracy. Refining the models, both qualitatively and quantitatively, using pilot testing data can further improve the accuracy.

The first six steps shown in Table 1 are focused on the qualitative layer of the design: the first decimal of accuracy. The next two steps describe the quantitative layer of the design, the second decimal of accuracy. Steps nine and ten refer to the refinement of the assessment design, which can improve both the qualitative and quantitative layers. Finally, step eleven refers to using the results of a stealth assessment (i.e., learners' competency estimates) for adaptive purposes (e.g., tailored learning supports per learner or adaptive challenges). This chapter revisits the first ten steps with a focus on maximizing the validity and reliability of the assessment to at least two decimal points of accuracy, drawing on experiences with the development of the game *Physics Playground* to illustrate the design considerations.

The Qualitative Design: How to Get the First Decimal Right

The qualitative steps of the stealth assessment design process focus on building the ECD domain model. Filling out the details of the domain model documents the central validity argument of the assessment. Information herein suggests that the learner does (or does not) have the target competencies because their work exhibited (or did not exhibit) certain observable features, and the assessment provides sufficient evidence of those competencies to be useful for the specified purpose.

The following discussion suggests that the design is a waterfall process, moving from one step to the next. However, in practice it is much more iterative, as later steps will uncover fuzzy places in the previous design which need better specification. Thus, although the sequence presented is Step 1, 2 and 3, the actual sequence is more like 1, 2, 1 again, 2 again, 3, 1 again, and so on.

Step 1. The Competency Model (CM)

The first step in any assessment design is identifying the target competencies and population. It is not sufficient to define what a competent person looks like, but the CM must also describe the typical paths learners take to reach competency. For this reason, Almond, Hernandez, and Turner (2021) refer to the CM as a *skill map;* the assessment must be able to locate the learner on their journey from base camp to the top of the mountain.

A key question to ask subject matter experts assisting in the design of the CM is: *What are expected differences in observable properties of learner work—especially for those close to and distal from the goal state*? It is helpful to think of the competency variables as ordered categorical variables at this stage, and have the experts apply *Russell's Rule*, which states that there should be at least one claim that can be made about learners in a higher level that cannot be made about learners at a lower level.

To produce a high-quality CM, assessment designers should (a) conduct an in-depth literature review about the targeted competency, and (b) consult with domain experts who know the competency and its pedagogy well (Step 1 in Table 1). When designing the stealth assessment of physics understanding in *PP*, we first identified standards in the Next Generation Science Standards (NGSS) relevant to the game and the target grade levels. To operationalize the competencies, we asked two experts to identify game behaviors they expected from learners who were high in the competency that they did not expect from learners low in the competency. In many cases, we could see those behaviors manifest within existing game levels. In other cases, new kinds of game levels were needed to capture evidence of the competencies.

Once the individual competencies were defined, we asked the experts to define the relationships among the variables. These relationships are generally of two types. The first reflects the natural hierarchy of the field of consideration. For instance, from a hierarchical perspective, Newton's laws are *part of* the larger concept of Force and Motion which in turn is *a part of* the larger concept of conceptual physics understanding. Not all relationships are hierarchical. For example, Properties of Momentum is a prerequisite to Conservation of Momentum because it is difficult to understand the conservation of momentum if momentum itself is not well understood. Other types of relationships may concern the way the skills are ordered in the curriculum; for example, students typically learn about force and motion before learning about energy, so there is a correlation between the two competency variables. The Bayesian network in Figure 12.2 (still qualitative at this stage) was our physics experts' best expression of the relationship among the targeted competency variables. To encourage the experts to discuss these relationships more fully, the design team presented them with two alternatives for Figure 12.2. The final figure captures the features that the experts thought best represented the space of learner competencies (Almond et al., 2017).

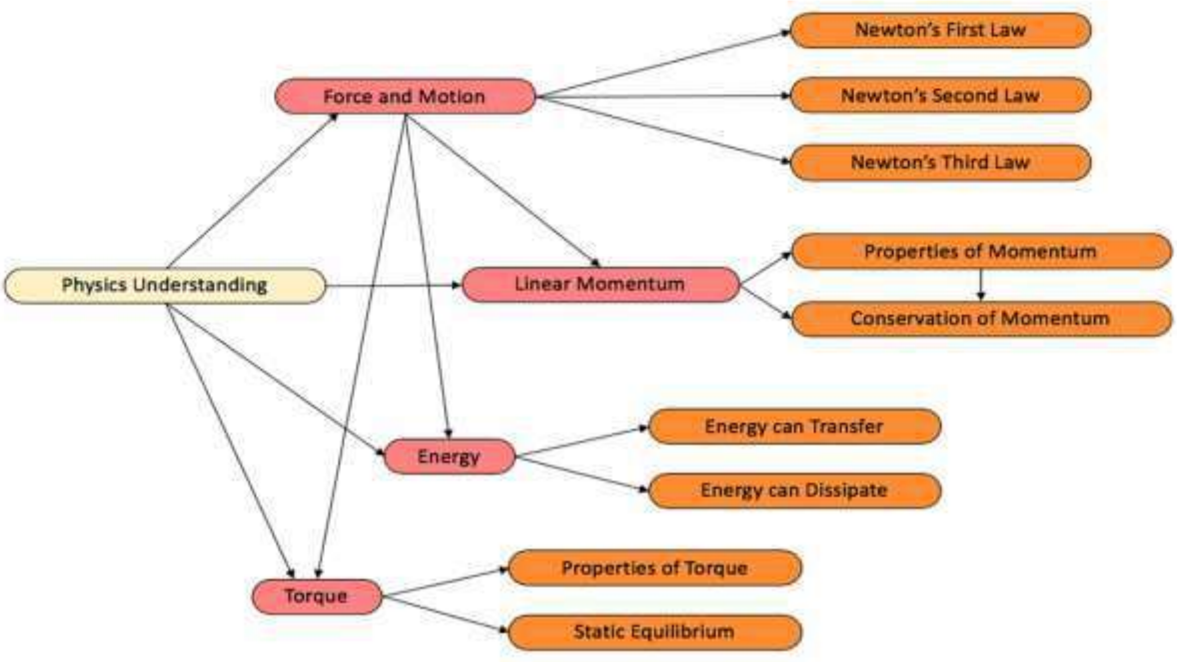

Figure 12.2: Physics understanding competency models.

Step 2. Selecting vs. Designing the Right Game

The second step in designing and developing a stealth assessment is to either choose an existing game or create one from scratch (see Table 1). Each of these options constrains the design and development process (see Smith et al., in press from this book for a detailed discussion on those constraints). For a game to serve as a stealth assessment, the actions required to distinguish between levels of competency must be a natural part of the mechanics. For example, open-ended, sandbox games such as *Minecraft* allow learners to express creativity through gameplay. In contrast, puzzle-based games such as *Plant vs. Zombies* are more suitable for assessing problem-solving skills (see Rahimi & Shute, 2021 for a systematic review on this topic).

When creating a game from scratch, choosing the right game genre, narrative, and game mechanics are crucial. For example, *Physics Playground*, a puzzle-based game equipped with a physics engine demonstrating the laws of physics, was specifically developed to assess middle school learners’ physics understanding. In *Sketching* levels (Figure 12.3a), learners can draw objects, create simple machines, and employ physics laws all toward reaching the goal of each level (i.e., to move a green ball so that it hits a red balloon). In *Manipulation* levels (Figure 12.3b), learners can adjust some sliders (altering the mass of the ball, gravity, and air resistance) or manipulate objects that can exert a force (i.e., a puffer or a blower) to solve the levels, similar to a simulation. An advantage of developing a new game instead of using an existing one is that specific capabilities can be added to the game to match evidentiary needs. In *PP,* the experts identified that understanding how the mass of the ball affects its interaction with other game objects was critical to understanding force and motion, and energy. Thus, the capability to manipulate the mass of the ball was added to certain game levels.

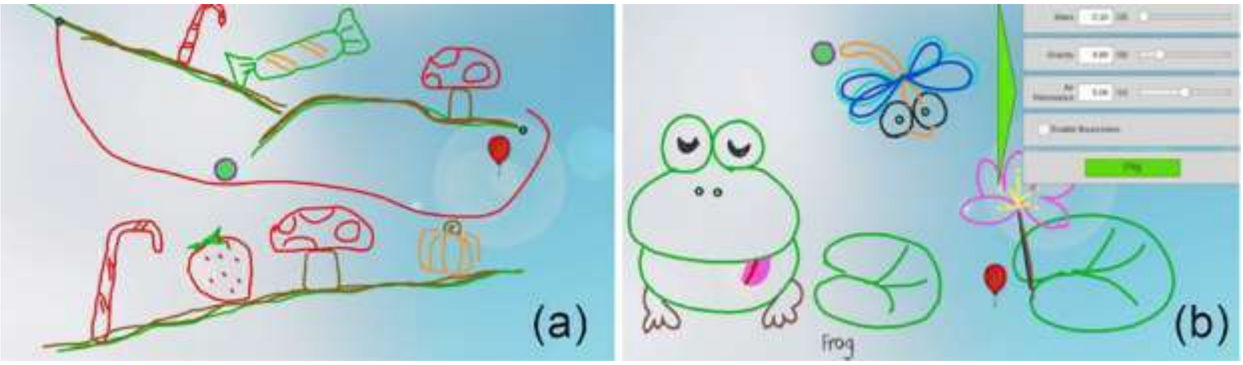

Figure 12.3: A Sketching level (a) and a Manipulation level (b) in PP.

Another example of a game designed from scratch to measure and support a competency is *Engage* (Min et al., 2020). *Engage* is a 3D game developed to teach computer coding (i.e., loops, conditions, variables) in a block-based manner. Learners need to progress through the game, unlock doors, and move their character (avatar), using block-based coding. The challenges of the game increase as the learner moves from one room to another. Again, the game designers of *Engage* first thought about *what game genre, narrative, and mechanics can be suitable for* *assessing and supporting programming or coding skills.* Then, the game developers started developing the game and called it *Engage*.

Apart from the affordances that a game can provide for assessing a particular competency, stealth-assessment designers should consider their target audience's reaction. Themes and/or graphics which strongly appeal to one demographic group may not appeal to other groups, or worse, may be culturally insensitive, which would damage the stealth assessment's fairness and hence validity. Feedback on the game design from designers and reviewers from various backgrounds, ethnicities, genders, and levels of education can spot such issues which can be overlooked by an insufficiently diverse design team. Seeking feedback both formally (via focus groups) and informally (from colleagues) can help here. However, the most essential is that projects' leadership must produce a climate where reviewers feel comfortable discussing potentially sensitive issues. A climate where comments about potentially sensitive issues are welcome and respected.

Another critical part of game design is the difficulty of the challenges. According to Csikszentmihalyi's flow theory (1990), there is an optimal difficulty at which flow (and hence optimal focus on the competencies measured by the game) will be induced. If the game is too hard or too easy for the target population, the learning and information gathered from the game will be sub-optimal. A complication here is that the word *difficulty* means different things in game design and psychometrics. Game difficulty is any game-related factor which makes the learner less likely to solve the game challenge (e.g., more obstacles between the ball and the balloon in *PP*); in psychometrics, difficulty generally refers to the amount of the target competencies required to solve the problem (e.g., a level whose primary concept is Newton's third law should be more difficult than a level with Newton's first law as its primary concept). In addition, many games rely on motor skills (e.g., pressing a key at the right time) to solve the challenge, but for an educational game, psychomotor skills would produce construct irrelevant variance in the measurement. Thus, alignment between the scoring mechanism of the game and the competencies to be assessed is key to getting reliable measurement.

Step 3. Evidence Models (EM)

Evidence models serve as the bridge between the competencies in the competency model and the task model (i.e., task work product). The observables are the pylons that support the bridge. Step 3 in Table 1 involves identifying such observables (or indicators), which represent the things that a person does (or says) in the game that provides evidence for some parts of the CM. For example, consider the evidentiary statement, "If the learner realizes they need to add energy to the ball to get it to the target (which is higher than the ball), then they will draw a springboard or a lever to add the energy." In this case, the competency in question is the understanding of potential energy, the task is a sketching level with the target (the balloon) higher than the starting position of the ball, and the observable is whether or not the learner drew a lever or a springboard.

At this stage, the observables (which are associated with tasks or task families) must be linked to the competencies they were intended to measure. This is often done with a *Q*-matrix—a matrix where rows correspond to observables and the columns to competencies. The entry to a cell in the matrix is 1 if that observable provides evidence for that competency, and 0 otherwise. Eventually (Step 7 in Table 1), somebody will need to produce computer code (or a human coding system) that will identify the exact value of the observable; defining observables precisely is required to provide specifications for computer software or human raters. Identifying a complete and informative set of observables is key to good assessment design. The stronger the evidence provided by the observables, the higher the information content of the assessment, and hence its reliability and validity. Note that the process of identifying the observables began when they were used to characterize the competencies but is now extended to cover all of the challenges within the game.

While designing the latest version of *PP*, the design team assigned a primary and a secondary competency to each game level. These competencies were chosen from the nine low-level competencies in the CM (the orange nodes in Figure 12.2), and assignments were reviewed by the physics experts. Looking at the numbers of entries in each column of the *Q*-matrix provides information about the amount of potential information for each competency. In the initial collection of levels, there were insufficient game levels targeting Newton's third law, so the design team created new game levels to fill this gap and ensure breadth of content coverage.

Although the final specifications of the observables (and the evidence rules that define them) will be established later, at this point in the design they should be at least tentatively identified. Identifying some of the observables in *PP* were straightforward. For instance, in terms of level solution, learners could earn a gold coin for an efficient solution (if the solution was conducted under a predefined condition such as using less than X number of objects in solving level Y), a silver coin for an inefficient solution (when the gold condition was not met but the learner solved the level), or no coin at all (when the learner quit a level without solving it). The amount of time spent on each level was also an obvious observable (although this proved to have little evidentiary value in pilot tests). For sketching levels, if the learner's drawing was an attempt at making a simple machine (i.e., a ramp, lever, springboard or pendulum), it provided evidence of physics understanding. This was challenging because extensive information from the physics engine was necessary to classify the learner drawings. That is, we required a classification system to be added to the game engine so it could log information about which simple machines were created and used. For example, to identify a pendulum in *PP*, the game engine checks if the object (e.g., a drawn line which might be a pendulum) touched the ball, if the object has a pin (which affixes the object to another object and is necessary for a pendulum), if the object rotated more than 20 degrees, and if the object has non-zero rotational velocity. If all the mentioned conditions are true, the game engine identifies the object as a pendulum.

The process of choosing appropriate observables when designing a stealth assessment requires extensive discussion among the design team (i.e., assessment designers, game developers, game-based learning researchers, subject matter experts, and developers responsible for the scoring systems). An alternative process is to use AI-based methods to automatically identify appropriate observables (e.g., Min et al., 2020). However, many machine learning algorithms are *black-boxes.* That is, machine learning models can identify a relationship (correlation) between a certain observation and a target competency, but cannot offer any explanation as to why that relation would hold. ECD, in contrast, is a *glass-box* method, thus intentionally transparent. The chain of reasoning that links an observable outcome variable and a latent competency variable is documented as part of the process, providing a strong construct validity argument. Machine learning techniques may be used at the cost of a weaker validity argument, but may be used to supplement experts' opinions.

Step 4. Task Models (TM)

In *PP,* tasks correspond to the game levels. *Task models* are descriptions of collections of tasks with an emphasis on the features that game designers can manipulate. Almond et al. (2014)

describe the various roles these features can play. Of particular importance are the features the affect the difficulty or evidential focus of the task. For example, if the balloon in *PP* is higher than the ball, then the learner must find a way to add energy to the ball, to counter gravitational forces. Designers must play particular attention to the features that affect both psychometric difficulty and game difficulty. While in more conventional assessments the goal is to maximize psychometric difficulty while minimizing game difficulty, in a stealth assessment the game difficulty may add to the challenge in a way that helps maintain a state of flow. As flow resides at the sweet spot between a learner's current ability and the game level's current difficulty. In *PP*, game designers coded each game level giving it a ranking on both the physics and game difficulty. Thinking about the game levels' difficulty using both physics and game difficulty helped us order the levels by using an incremental composite level of difficulty (i.e., the sum of the two difficulty scores per level).

There are two other key roles for task variables, the first being that they determine which competencies are relevant. For example, to provide evidence about understanding Newton's second law ($F = m \times a$), the learner needs to be able to manipulate an object's mass (e.g., in *PP* the manipulation levels allow for the manipulation of mass). This type of task variable is used to match the task with the appropriate evidence model. The second role for task variables is to ensure that all facets of the competency are explored. For example, in *PP*'s manipulation levels (see Figure 12.3b for an example), forces could come from the gravity slider or from puffers and blowers. A given level could choose any of those sources of forces, but the final collection of tasks should include multiple examples of each source.

Steps 5 and 6. The Q-matrix and the Assembly Model

A game consists of a series of challenges (e.g., levels in *PP,* but these could be expressed in different ways and in different game types) and rules for forming the sequence of challenges for the learners. These series of challenges and sequencing rules determine the narrative structure of the game, but also the quantity of evidence available for estimating learners' competencies. In particular, the reliability with which the stealth assessment can measure each competency is largely determined by how many tasks provide evidence for each competency. To maximize validity, the collection of tasks must span the breadth and depth of the competency. These psychometric needs must be balanced with the time constraints and the need for the game to hold learners' interest.

One of the tools for managing the collection of levels is the *Q*-Matrix (steps 5) in Table 1). Table 2 shows a portion of the *Q*-Matrix used for *PP*. In this *Q*-matrix, a 1 denotes the primary competency and 2 a secondary competency (e.g., in Table 2, Level 1 has NFL as its primary competency and NSL as its secondary competency). The *Q*-matrix included columns indicating the game mechanics difficulty (GM) and physics understanding difficulty (PU) of each game level. The composite difficulty score was the sum of GM and PU (discussed in more detail later). Counting the number of entries in each column provides a quick estimate of the amount of evidence available for the corresponding competency ensuring a complete coverage of the competency model. Including the difficulty information in the *Q-matrix* allows the project manager(s) to ensure that the tasks span the relevant depths of competencies of interest and that the game will not be too challenging and hence not provide useful information.

Table 2
Example of the Q-Matrix

	NFL	NSL	NTL	POM	COM	ECT	ECD	POT	Equil.	GM	PU	Comp.
Level 1	1	2	0	0	0	0	0	0	0	2	3	5
Level 2	0	0	0	0	0	1	0	2	0	3	5	8
…	…	…	…	…	…	…	…	…	…	…	…	…
Level n	1	2	0	0	0	0	0	0	0	2	2	4

Note. NFL = Newton First Law; NSL = Newton Second Law; NTL = Newton Third Law; POM = Properties of Momentum; COM = Conservation of Momentum; ECT = Energy Can Transfer; ECD = Energy Can Dissipate; POT = Properties of Torque; Equil. = Statis Equilibrium; GM = Game Mechanics difficulty; PU = Physics Understanding Difficulty; Comp. = Composite difficulty.

In a typical stealth assessment development process, the full *Q*-matrix defines the complete collection of all available tasks or game levels. Then, for a given use of the stealth assessment (e.g., assessment of two rather than all nine competencies in *PP*), the game designers pick a subset of the tasks. In that case, it is the selected subset *Q*-matrix that must be properly balanced; that is, have complete coverage and enough levels per competency to meet the needs of the assessment.

In addition, using the assembly model (AM) the assessment and game designers must specify how the game levels are sequenced. Three common sequences include: (a) *linear sequencing* in which all learners see the game tasks in the same sequence; (b) *random selection* where the game tasks are drawn (without replacement) from a pool of potential tasks; and (c) *free choice* where the learner chooses from a menu of possible tasks. Combinations of these strategies are also possible, often accomplished by grouping tasks in stages. For example, there could be a linear sequence of stages and free choice within each stage. Although free choice seems like it would be valued by learners, in most cases, the learners do not have enough information about the tasks to make informed choices and default to the order in which the tasks are presented to them.

Another way to present game levels to learners is via adaptive sequencing or selection. Here the next game level is chosen by the system based on the current estimate of learner ability from the stealth assessment. There seems to be some evidence that sequences chosen to optimize information about learners' ability also optimize learners' learning. The current version of *PP* has linear, free choice, and adaptive sequencing options. The adaptive option grouped game levels into stages corresponding to the nine lower-level competency variables (the primary classification in the *Q*-matrix). The stages were then sequenced with the advice from the physics experts. Learners would be given game levels from the current stage until (a) sufficient evidence was gathered that the learner had mastery of that competency, (b) sufficient evidence of lack-of-mastery was gathered that it would be best to move to a different stage, or (c) the game ran out of levels in that stage and the learner moves to a new stage. If the learner completed all of the stages, they reached an end-game where the system would give them unsolved levels from any stage or they would be encouraged to go back and earn gold coins for levels where they had only earned a silver coin.

Finally, the assembly model must also determine when it is time to stop. Often (particularly in experimental settings) this is determined by task duration. Another common termination condition is completing all available levels. If the game is adaptive, an adaptive selection system could use the amount of evidence gathered about the learner to terminate the game as well. It is important to ensure that the learners have enough time to provide sufficient evidence. For instance, if the learners only have time to complete 3 of 25 planned levels, the estimates are unlikely to be sufficiently reliable for the intended purpose and scores should not be reported (at least without strong warnings). Note that when the purpose of a stealth assessment is summative, an end point is relevant; however, when the purpose of stealth assessment is formative (for learning), determining a stop point may not be relevant.

Quantitative Model: How to Get the Second Decimal Right

Getting the collection of tasks right is the first decimal place of accuracy because if the tasks do not provide the right evidence, no amount of statistical modeling can save the assessment. However, the statistical model can only improve the reliability and validity of the assessment within the limits set by defining the competencies, tasks and evidence. Therefore, statistical modeling can only provide the second decimal of accuracy. Designers can choose between a data-driven approach (e.g., either using statistical models like item response theory, regression, and factor analysis, or machine learning approaches like neural networks), or an expert-driven approach (e.g., such as number right scoring or expert systems). Data-driven approaches have two disadvantages: (1) they require a lot of data (in many cases, thousands of subjects), and (2) they can be difficult to explain, taking advantage of correlations between competencies and observations without explaining their causes. Bayesian networks (BNs) share the benefits of expert-driven and data-driven approaches. Expert opinion can be used to form initial models, and these models can be updated and validated when data are available for BNs' training. Furthermore, BNs provide an explanatory structure to allow them to be reviewed by outside experts (i.e., using a graphical representation of the network).

The specification for scoring machinery (steps 7 and 8 in Table 1) is mostly in the EM with some important contributions from the other ECD models. The EM has two parts: (1) the *evidence rules,* which along with the specifications of the work products (event logs) from the

TM, provide the specifications for the evidence identification process; and (2) the *statistical model*, which along with the description of the population distribution of the competencies (CM), provide the specifications for the evidence accumulation process.

Implementing the Rules of Evidence

Step 7 of the ten-step process (see Table 1) is about clearly defining the observables. These observables will be the output of the evidence identification process and the input to the evidence accumulation process. For example, if *duration to solve a level in a game* is one of the observables, one can score this observable as 0 if the duration is beyond the acceptable range (showing that the student struggled), 1 if the duration is between the acceptable range (showing that the student is doing good), and 2 if the duration is below the acceptable rage (showing mastery and excellence). Note that duration might not always be a good observable and it depends on the competency and the game. Once this observable is scored (evidence identification), it is ready to be used by the evidence accumulation process which includes the statistical modeling. Observables are often ordered, categorical variables; where the first step is often to write a *rubric,* a set of descriptions for each possible value of the observable. It is possible for human raters to be involved in the evidence identification process, although usually scores are needed more quickly than human raters can provide them. Even so, human scored examples of learner work may be important in training machine learning approaches or testing algorithmic scoring approaches.

PP used an approach that combined elements of a finite state machine with elements of a rule-based system (Almond et al, 2020). While the goal was to allow a simpler translation from natural language to computer instructions, in practice only people with coding experience could easily write the rules. According to our experience, the coordination between the team coding the evidence identification and presentation processes if crucial for a smooth implementation process (e.g., the location of various kinds of information in the log files needs to be clearly documented).

An alternative approach at this stage is to use machine learning algorithms to identify “features” of the output. Again, a lot of pilot data is needed for the data-driven approaches, and work on the statistical model cannot really begin before the observables are identified. In *PP*, we used some data driven techniques in revising the models after all the data were available; particularly setting cut points for turning continuous measurements into ordered categories based on quantiles of the observed quantities.

Choosing the Right Statistical Modeling Approach

De Klerk, Veldkamp, and Eggen (2015) conducted a systematic review of the statistical modeling approaches used in game-based assessment. The results of their review included various modeling approaches: confirmatory factor analysis, educational data mining techniques, item response theory, artificial neural networks, cluster analysis, and Bayesian networks (BNs). De Klerk and colleagues concluded that the most frequently applied modeling approach to analyze performance-based data in games and simulations was BNs.

There are several advantages to using BNs for statistical modeling of learners’ performance estimates. Almond et al. (2015) provide four main points: (1) BNs provide an easy-to-view graphical representation of the CM (direct and indirect relationships among variables); (2) BNs can initially be built using expert opinion and then refined using Bayesian learning as more data become available, which is essential when designing a stealth assessment; (3)

Updating BNs happens quickly, as observables come directly from the evidence identification process thus providing real-time competency estimates which can be used for adaptive task selection and learning support presentation; and (4) Enhancements to BN software permit large and flexible networks with as many variables and connections among them as wanted. For these reasons, we recommend choosing BNs for any stealth assessment design.

Graphical Structure of BNs

Again, designing a stealth assessment needs a team with diverse backgrounds and expertise. The measurement team, in charge of creating the statistical model and BNs, should communicate crucial pieces of information to other team members in a way that they understand—especially the subject matter experts—for getting their feedback. For instance, at least two versions of the general graphic BN model should be created and shown to the experts for feedback and model selection.

To create the general BN in *PP*, we first conducted a review of relevant research and consulted with our physics experts to identify the most important variables (i.e., the competencies and their sub-facets related to physics understanding). Then, using the experts' input coupled with our review, we created two graphical models using *Netica* (Norsys, 2021), and showed them to our experts to edit and then choose one. This stage includes no technical prerequisite knowledge about BNs. Showing at least two candidate models is helpful, as giving only one model to evaluate may limit deep thinking about the variable's relationship (e.g., pedagogical order of physics concepts). Moreover, when showing two models to the experts, it is crucial to talk about the conditional independence of each variable. For example, we asked our physics experts if Newton's laws should be conditionally independent or dependent on force and motion (i.e., if understanding force and motion is dependent or independent of Newton's laws of motion). The lines from force and motion to Newton's laws in Figure 12.2 shows that understanding force and motion is conditionally dependent given Newton's laws. Once the BN model's graphical structure is finalized, the measurement team can start working on another critical aspect of the BNs—the conditional probability tables (CPTs). To specify the CPTs, the measurement team needs to specify the *type* of relationship existing among CM variables (i.e., conjunctive, disjunctive, or compensatory), difficulty, and discrimination values for each task in the network.

Conjunctive, Disjunctive, and Compensatory Models

In traditional statistical modeling approaches (e.g., simple Item Response Theory), only one parent node is allowed for the higher-level competency (e.g., reading or writing on an English test). To solve a particular task, only one skill is needed (e.g., writing skill is needed for doing well on an essay). However, BNs allow for multiple parent nodes at competency level (aka, proficiency level; Almond et al., 2015). Thus, when BNs are used for statistical modeling, it is possible to have a task that requires more than one skill. When a task requires more than one skill, the relationship among those skills should be specified. For instance, does the learner need both skills at the same time to be able to perform a task (i.e., solve a game level)? Do either of the skills suffice for a successful performance? There are three ways that the parents of a node in a BN (i.e., an observable linked to more than one competency variable) can relate to each other: conjunctive, disjunctive, compensatory relations. First, a *conjunctive* relationship indicates that all skills (i.e., parent nodes) are needed for a high probability of a successful performance in the task at hand (e.g., when both reading and writing skills are needed to write an essay). The conjunctive relationship of the parent nodes are understood as the "and" operator. Second, the *disjunctive* relationship indicates that the presence of one of the skills associated with the task at hand is enough for successful performance. The disjunctive relationship of the parent nodes may be viewed as the “or” operator, and disjunctive tasks can be solved with alternative solutions. Third, the *compensatory* relationship indicates that being high on one skill compensates for being low or medium on the other relevant skills. In this case, the effect of the skills is additive and the probability of a successful performance can be computed as the sum of the skills.

In a BN, each observable needs to have a conditional probability table (CPT). Before computing each CPT, assessment designers need to specify these relationships. This means that a BN can have multiple observables that have more than one type of modeling (e.g., a mixture of conjunctive and disjunctive relationships for different observables). In other words, any of the modeling approaches discussed above can be included in a BN, making the model more complex, yet more accurate. Choosing the most theoretically meaningful relationship among the required skills (i.e., the appropriate modeling) for an observable has ramifications for generating an accurate CPT. Thus again, consultation with subject matter experts is warranted. Getting these relationships wrong can lead to inaccurate estimates of learners’ competency levels. Next, we discuss the difficulty and discrimination of tasks or observables that must be included in each CPT specification.

Difficulty and Discrimination

Generally, in the world of assessment and measurement, item difficulty directly relates to item discrimination. An item which is too difficult (i.e., a small percentage of learners can answer it correctly) or too easy (i.e., a large percentage of learners can answer it correctly) is an item with low discriminating power (i.e., discriminating between learners who are high and low on the targeted competency). Item discrimination indicates to what extent correctly responding to the item relates to general success on the whole test. The items (observables) with good discrimination tend to correlate well with the total score of the learner on the competency of interest—also known as item-to-total correlation (DeVellis, 2006). Therefore, assessment designers need to conduct pilot tests and evaluate their items in terms of difficulty and discrimination parameters.

The same concept can be applied to stealth assessment. Game levels and any observable can be seen as test items. Each game level has a difficulty index. Each level's difficulty in an educational game can be a function of the composite of two difficulty indices: game mechanics' difficulty and the concept difficulty. We have investigated the effects of these two types of difficulty on learners' persistence (Rahimi et al., 2021) and frustration (Karumbaiah et al., 2018) in *PP*. In both cases, concept difficulty showed more predictive power in predicting learners' persistence and frustration. Identifying the difficulty of each level needs rigorous qualitative and quantitative approaches. For example, to derive our two difficulty indices for each game level in *PP* (i.e., related to game mechanics and physics understanding), we first brainstormed the criteria for game mechanics difficulty (GM). The following are the main criteria: relative position of the ball to the balloon (e.g., if the ball was above the balloon, it was easier to solve the level than when the ball is below the balloon, as one need to defy gravity); the number of obstacles between the ball and the balloon (e.g., a level with no obstacles between the ball and the balloon was easier to solve than a level with several obstacles); the degree of precision for the level's solution (e.g., if the balloon was trapped in a pipe and the pipe's exit hole was small, it was more difficult to solve the level than when the balloon was on a flat surface); and if the level's name provided any hint for the solution (e.g., *Perfect Pendulum*). Each of these criteria would receive a score (some 0 or 1; others 0 to 2).

For physics understanding difficulty (PU), we consulted with our physics experts. We asked them to identify each game level's primary and secondary concepts. They provided the criteria for identifying the difficulty of each level based on the primary and secondary concept(s) or competencies associated to that level. For example, if the primary and secondary physics concepts of a level arose from the same parent node in the CM, the game level is easier than a game level with two concepts from two different parent nodes. The other criteria for PU, suggested by our experts, was the order of the topics. In general, our physics experts rated the force and motion concepts (i.e., Newton's laws) easier than the torque concepts (e.g., equilibrium). Once the criteria (rubric) for rating game levels' GM and PU was established, two raters independently rated the difficulty of each game level. Then, the two raters resolved their disagreements and came to a 100% agreement. Finally, we summed GM and PU to get a composite difficulty score for each game level. Table 2 shows examples of game levels with various difficulty values.

There are other methods to compute the difficulty indices of each game level or observable in a game environment. For example, one can average the time it takes for learners to solve a game level, positing that the more time it takes to solve a game level, the more difficult that level. We tested this hypothesis. The average time it takes to solve game levels in *PP* is significantly correlated with our composite difficulty index (r = .60). From a game design perspective, when including various game levels with various difficulties in a game, it is a recommended to array the levels from easy to difficult. It is important to note again that in stealth assessment we treat each observable just like a test item. That is, each observable has its own difficulty and discrimination values.

One can think about the difficulty and discrimination in terms of the intercept and the slope of a line, respectively. Consider a graph with learner proficiency level on Skill A on the x axis, and the probability of the learners correctly responding to that item on the y axis. If the task (i.e., a game level or any observable) is highly difficult, the intercept is high, and the slope probably becomes very steep. Alternatively, an easy task has a low intercept and again steep slope. In both cases, we have low discriminating power. An ideal situation occurs when an item has a medium to high difficulty and a reasonably steep slope. When specifying the CPT, measurement team members need to think carefully about parameters, consult with the experts, and choose the difficulty and discrimination indices carefully, per observable. These two metrics directly impact the CPTs—discussed next.

Conditional Probability Tables

Setting up the conditional probability tables (CPTs) of the Bayes net should include input from the content experts. For example, when creating the CPTs for *PP*, our subject-matter experts were physicists, and they were very familiar with our targeted competencies/concepts. However, even when the experts are familiar with those concepts, the stealth assessment designers need to come up with understandable statements that can help the experts think in probabilistic terms.

To help facilitate setting up CPTs, the measurement team needs to first compute initial CPTs. While this process can be done with software (e.g., Netica), we recommend using the open-source *Peanut* package (Almond, 2021), which provides functions that compute the conditional probability tables in the R language (R Core Team, 2021). To read more about this process, see Almond et al. (2017). If the conditional probability distributions for all nodes in the BN model are expressed using R statements, the experts might not understand the language. Therefore, the measurement team needs to translate those statements into natural language. Below, in Table 3, you can see an example from *PP*.

Table 3

An Example of Translating the CPT to Natural Language (from Almond et al., CITE)

R Code for the Force and Motion CPT	Translated CPT for Force and Motion
eng <- PP.High$Energy PnodeRules(eng) <- "Compensatory" PnodeLink(eng) <- "normalLink" PnodeLinkScale(eng) <- sqrt(.2) PnodeLnAlphas(eng) <- log(c(Physics=sqrt(.7),ForceAndMotion= sqrt(.9))) PnodeBetas(eng) <- 0	*Force and Motion: its parent is physics only. We are setting a regression of force and motion on physics understanding.* *Link scale parameter only gives us R-squared, which is the percent of the explained variance by the predictors on the outcome variable. The value of R-squared is 0.8 (1 – the variance unexplained which is 0.2).* *The shift of about half of the standard deviation (0.5) up towards more people having the skill. The shift is telling us a person who is medium on the parent variable is going to be somewhere about halfway between medium and high. Most of the weights were split between medium and high.*

Other than providing the translated version of a CPT, the computed CPT can also be generated and shared with the experts for input. For example, the CPT of Force and Motion is shown in Table 4. Note that Force and Motion is dependent on Physics (i.e., general physics understanding), and the probability distributions of Force and Motion are computed given the probability distribution of Physics.

Table 4

The CPT for Force and Motion and Physics Generated by the Peanut package

		Force & Motion	
Physics	High	Medium	Low
High	0.98	0.02	0.00
Medium	0.56	0.42	0.02
Low	0.04	0.52	0.44

According to Table 4, almost all of the learners who are high on the overarching Physics variable will also be high on force and motion. Moreover, learners who are low on Physics tend to be divided between medium and low on Force and Motion. Giving these pieces of information to the physics experts and asking them, “Do these values make sense to you? If not, what changes would you make?” can help a lot with establishing the BN’s initial state at a more accurate level relevant to the learner population of interest. Note that the process of consulting with the experts on CPTs can be an iterative process rather than a one-time event. When the BNs start running with accurate CPTs, they will learn faster from the data coming in. Alternatively, starting the BNs at an inaccurate state can inflate the final stealth assessment estimates.

Evaluating the Validity, Reliability, and Fairness of a Stealth Assessment

In this section, we provide suggestions to evaluate the psychometric features of stealth assessments in terms of validity, reliability, and fairness (steps 8 and 9 in Table 1). These evaluation ideas generally apply to all types of statistical models used for assessment, including IRT and Bayes nets.

As discussed in the introduction, validity and reliability usually go hand in hand. The ECD framework that underlies stealth assessment serves to ensure the reliability and validity of the assessment starting from the conceptualization and design processes. One common practice is to check the convergent validity of a stealth assessment by correlating the results generated from gameplay with external established measures. For example, to validate the stealth assessment of learners' physics understanding, Shute et al. (2020) correlated the estimates from the game's BNs with validated, expert-created physics test scores. Results showed that the stealth assessment of physics understanding correlated with both pretest (r = .36, p < .01) and posttest (r = .40, p < .01). In another similar study, Shute and Rahimi (2021) reported that the stealth assessment of creativity in *PP* significantly correlated with the external measures of creativity ranging from .18 to .23. Other studies have reported a range of correlations from .40 to .67 (DiCerbo et al., 2016). Note that finding large correlations between stealth assessment estimates and traditional tests may not be expected (DiCerbo et al., 2016). That is, stealth assessment aims to assess learners' knowledge and skills on a deeper level (e.g., including the process data) compared to the traditional assessments (e.g., multiple-choice tests). Therefore, one should not consider small but significant correlations as evidence for lack of convergent validity of a stealth assessment; Significant, small to moderate correlations tell us we are going in the right direction (DiCerbo et al., 2016).

In terms of reliability, the split-half reliability method can be conducted (Shute et al., 2008) where a test is divided into two halves and the correlation between learners' scores on the two halves of the test is computed. Ideally, this correlation should be significant and positive to indicate a highly reliable test. In a game like *PP*, which is level-based, we can split levels into halves. Thus, researchers dealing with similar games should define what a meaningful chunk of the game is when doing split-half reliability checks. Issues arise when a game is not level-based, such as sandbox games (e.g., open-ended games such as *Minecraft*) or simulations (e.g., a flight simulator). It is critical to define the boundary of a task, context, or item in those cases. One idea is using cross-validation, where a portion of levels is used to predict the performance of other levels on estimating learners' competency. Cross-validation can be suitable, especially for predictive models, such as Bayes nets.

Traditional reliability coefficients can also be computed using the stealth assessment estimates. DiCerbo et al. (2016) discuss how Cronbach's alpha can be used to establish the reliability of game-based and performance-based assessments. They also indicate that Cronbach's alpha assumes that all items in an assessment load on a single construct. This assumption, if violated, can produce bias in assessment results. Therefore, other coefficients such as omega can be considered for assessments that examine more than one construct (e.g., stealth assessment).

Moreover, differential item functioning has been frequently used to statistically evaluate the fairness of assessments when Item Response Theory modeling was used (Holland & Wainer, 1993). One method to test the bias of an assessment between matched groups is called the Mantel-Haenszel test (Holland & Thayer, 1986). To compare the performance of two groups (the focal and reference groups), first the students are split into strata on the basis of their ability (this might be categories from the scoring model, or cuts on the number of game levels solved). Within each stratum, the probability of success for all score categories on the task should be the same. If the difference is larger than can be explained by random variation in performance (i.e., the Mantel-Haenszel test is significant), then the corresponding task should be investigated to see if it depends on information possessed by the reference group and not the focal group.

Specific to stealth assessment, Almond et al. (2015) proposed conducting differential task functioning and observable characteristic plots to evaluate fairness. Differential task functioning is intended to check whether the assessment behaves the same way for all subgroups of learners. Ideally, the evidence model should capture all performance differences among subgroups of learners. That is, conditioning on learners' competency, the performance elicited by a certain task should be the same across all subgroups of learners. Performance on tasks should be conditionally independent given the competency variables. For example, boys and girls with the same ability in physics should have the same performance.

An observable characteristic plot introduced by Almond et al. (2015) is a graphic representation used to detect possible bias impacting subpopulations. Such plots group learners based on their competency profiles and each group of learners has an estimated probability of their competency given their observed performance. The plots can serve multiple purposes: (a) researchers can plot observable variables against the competency variables to determine whether the observable variables provide relevant evidence to the associated competency variables; (b) researchers can compare two different models using the plot by adding a comparison value as another competency variable; and (c) the plots can compare focal and reference groups (e.g., boys vs. girls). Finally, before quantitative evaluation, fairness should also be considered during subject recruitment and data collection procedures. For example, researchers can try to recruit people who have various backgrounds, such as minorities, non-gamers, and learners with low prior-knowledge or low computer skills.

Conclusion

In this chapter, we defined the three important psychometrics features of an assessment – validity, reliability, and fairness. If assessment designers focus attention on enhancing these three aspects in any assessment (including stealth assessment), the quality of their assessment will increase. Alternatively, insufficient attention to these assessment qualities can lead to poor assessments with unexpected and undesirable consequences. For instance, if a stealth assessment is designed poorly, the estimates will likely be inaccurate. Then, if those estimates are used to provide relevant learning supports during gameplay, low-performing learners who need supports may not receive them, while high-performing learners who do not need supports may receive them. In this case, low-performing students can get frustrated and quit, while high-performing students may get frustrated due to the unnecessary and bothersome interruptions.

To ensure the estimates of a stealth assessment are accurate, stealth assessment designers need to use psychometrically sound frameworks to design their assessments (e.g., ECD). In this chapter, we grouped our recommendations into two categories: considerations to get the first and the second decimals of an estimate right, emphasizing the importance of the first decimal. This chapter represents just the tip of an iceberg on how to design high-quality stealth assessments using ECD. Please note that designing a stealth assessment is a complex yet promising task that needs to be done by a team of researchers—game designers, learning scientists, measurement and subject matter experts. We believe that the future of education will benefit from assessments such as stealth assessments that can improve how we learn. Thus, researchers and educators need to work on increasing the quality of such assessments.

References

Almond, R. G. (2020). Scoring of Streaming Data in Game-Based Assessments. In *Handbook of Automated Scoring*. Chapman and Hall/CRC.

Almond, R. G. (2021). *Peanut: Parameterized bayesian networks, abstract classes*. https://pluto.coe.fsu.edu/RNetica

Almond, R. G., Kim, Y. J., Velasquez, G., & Shute, V. J. (2014). How task features impact evidence from assessments embedded in simulations and games. *Measurement: Interdisciplinary Research and Perspectives*, *12*(1–2), 1–33. https://doi.org/10.1080/15366367.2014.910060

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian Networks in Educational Assessment*. Springer.

Almond, R., Steinberg, L., & Mislevy, R. (2002). Enhancing the Design and Delivery of Assessment Systems: A Four-Process Architecture. *Journal of Technology , Learning, and Assessment*, *1*, 1–64.

Almond, R., Tingir, S., Lu, X., Sun, C., & Rahimi, S. (2017, August 15). An elicitation tool for conditional probability tables (CPT) for Physics Playground. *Uncertainty in Artificial Intelligence*.

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

Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience*. Harper and Row.

de Klerk, S., Veldkamp, B. P., & Eggen, T. J. H. M. (2015). Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example. *Computers & Education*, *85*, 23–34. https://doi.org/10.1016/j.compedu.2014.12.020

DeVellis, R. F. (2006). Classical test theory. *Medical Care*, *44*(11), S50–S59.

DiCerbo, K. E., Shute, V., & Kim, Y. J. (2016). The future of assessment in technology-rich environments: Psychometric considerations. In M. J. Spector, B. B. Lockee, & M. D.

Childress (Eds.), *Learning, Design, and Technology* (pp. 1–21). Springer International Publishing. https://doi.org/10.1007/978-3-319-17727-4_66-1

Dorans, N. J., & Cook, L. L. (2016). *Fairness in Educational Assessment and Measurement*. Routledge.

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the ManTEL-Haenszel procedure. *ETS Research Report Series*, *1986*(2), i–24. https://doi.org/10.1002/j.2330-8516.1986.tb00186.x

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning* (pp. xv, 453). Lawrence Erlbaum Associates, Inc.

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

Karumbaiah, S., Rahimi, S., Baker, R. S., Shute, V., & D'Mello, S. K. (2018). Is student frustration in learning games more associated with game mechanics or conceptual understanding? In J. Kay, R. Luckin, M. Mavrikis, & K. Porayska-Pomsta (Eds.), *International Conference of Learning Sciences* (pp. 1–2).

Messick, S. (1994). *Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performancesas Scientific Inquiry into Score Meaning*. 33.

Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Smith, A., Wiebe, E., Boyer, K. E., & Lester, J. C. (2020). DeepStealth: Game-Based learning stealth assessment with deep neural networks. *IEEE Transactions on Learning Technologies*, *13*(2), 312–325. https://doi.org/10.1109/TLT.2019.2922356

Mislevy, R. J., Haertel, G., Cheng, B. H., Ructtinger, L., DeBarger, A., Murray, E., Rose, D., Gravel, J., Colker, A. M., Rutstein, D., & others. (2013). A “conditional” sense of fairness in assessment. *Educational Research and Evaluation*, *19*(2–3), 121–140.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research & Perspective*, *1*(1), 3–62. https://doi.org/10.1207/S15366359MEA0101_02

Norsys. (2021). *Norsys Software Corp. - Bayes Net Software*. https://www.norsys.com/

R Core Team. (2021). *R: The R project for statistical computing*. https://www.r-project.org/

Rahimi, S., & Shute, V. J. (2021). The effects of video games on creativity: A systematic review. In S. W. Russ, J. D. Hoffmann, & J. C. Kaufman (Eds.), *Handbook of lifespan development of creativity* (pp. 1–37). Cambridge University Press.

Rahimi, S., Shute, V., & Zhang, Q. (2021). The effects of game and student characteristics on persistence in educational games: A hierarchical linear modeling approach. *International Journal of Technology in Education and Science*, *5*(2), 141–165. https://doi.org/10.46328/ijtes.118

Shute, V., Almond, R., & Rahimi, S. (2019). *Physics Playground* (1.3) [Computer software]. https://pluto.coe.fsu.edu/ppteam/pp-links/

Shute, V. J. (2009). Simply Assessment. *International Journal of Learning and Media*, *1*(2), 1–11. https://doi.org/10.1162/ijlm.2009.0014

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., Hansen, E. G., & Almond, R. G. (2008). You Can't Fatten A Hog by Weighing It – Or Can You? Evaluating an Assessment for Learning System Called ACED. *International Journal of Artificial Intelligence in Education*, *18*(4), 289–316.

Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior*, *116*, 106647. https://doi.org/10.1016/j.chb.2020.106647

Shute, V., Lu, X., & Rahimi, S. (2021). Stealth assessment. In J. M. Spector (Ed.), *The Routledge Encyclopedia of Education* (p. 9). Taylor & Francis group.

Shute, V., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C.-P., Kuba, R., Liu, Z., Yang, X., & Sun, C. (2020). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity and learning supports in educational games. *Journal of Computer Assisted Learning*, *37*(1). https://doi.org/10.1111/jcal.12473

Smith, G., Shute, V. J., Rahimi, S., Kuba, R., & Dai, C.-P. (in press). Stealth assessment and digital learning game design. In M. P. McCreery & S. K. Krach (Eds.), *Games as Stealth Assessments*. DOI press.