

Competency model development: The backbone of successful stealth assessments

Seyedahmad Rahimi¹  | Russell Almond² | Andrea Ramírez-Salgado¹ |
Christine Wusylko¹ | Lauren Weisberg¹ | Yukyeong Song¹ | Jie Lu¹ |
Ted Myers¹ | Bowen Wang¹ | Xiaomaon Wang¹ | Marc Francois¹ |
Jennifer Moses¹ | Eric Wright¹ 

¹School of Teaching and Learning, University of Florida, Gainesville, USA

²Educational Psychology and Learning Systems, Florida State University, Tallahassee, USA

Correspondence

Seyedahmad Rahimi, School of Teaching and Learning, College of Education, University of Florida, 2821 Norman Hall, PO Box 117048, Gainesville, FL 32611, USA.
Email: srahimi@ufl.edu

Abstract

Background: Stealth assessment is a learning analytics method, which leverages the collection and analysis of learners' interaction data to make real-time inferences about their learning. Employed in digital learning environments, stealth assessment helps researchers, educators, and teachers evaluate learners' competencies and customize the learning experience to their specific needs. This adaptability is closely intertwined with theories related to learning, engagement, and motivation. The foundation of stealth assessment rests on evidence-centered design (ECD), consisting of four core models: the Competency Model (CM), Evidence Model, Task Model, and Assembly Model.

Objective: The first step in designing a stealth assessment entails producing operational definitions of the constructs to be assessed. The CM establishes a framework of latent variables representing the target constructs, as well as their interrelations. When developing the CM, assessment designers must produce clear descriptions of the claims associated with the latent variables and their states, as well as sketch out how the competencies can be measured using assessment tasks. As the designers elaborate on the assessment model, the CM definitions need to be revisited to make sure they work with the scope and constraints of the assessment. Although this is the first step, problems at this stage may result in an assessment that does not meet the intended purpose. The objective of this paper is to elucidate the necessary steps for CM development and to highlight potential challenges in the process, along with strategies for addressing them, particularly for designers without much formal assessment experience.

Method: This paper is a methodological exposition, showcasing five examples of CM development. Specifically, we conducted a qualitative retrospective analysis of the CM development procedure, wherein participants unfamiliar with ECD applied the framework and showcased their work. In a stealth assessment course, four groups of students (novice stealth assessment designers) engaged in developing stealth

assessments for challenging-to-measure constructs across four distinct projects. During their CM development process, we observed various activities to pinpoint areas of difficulty.

Results: This paper presents five illustrative examples, including one for assessing physics understanding and four for the development of CMs for four complex competencies: (1) systems thinking, (2) online information credibility evaluation, (3) computational thinking, and (4) collaborative creativity. Each example represents a case in CM development, offering valuable insights.

Conclusion: The paper concludes by discussing several guidelines derived from the examples discussed. Emphasizing the importance of dedicating ample time to fine-tune CMs can significantly enhance the accuracy of assessments related to learners' knowledge and skills. It underscores the significance of qualitative phases in crafting comprehensive stealth assessments, such as CMs, alongside the quantitative statistical modeling and technical aspects of these assessments.

KEYWORDS

assessment for learning, competency model, digital games, evidence centered design, learning analytics, stealth assessment

1 | INTRODUCTION

Teach. Stop. Test. Stop. Teach. Stop. Test. Stop...

This is the usual pattern in educational systems. Much valuable time and energy is wasted in this interrupted format. Those “Stop. Test. Stop.” times could be used for learning (i.e., deepening learners' knowledge and skills). Moreover, much time goes into test preparation and grading. Assessments separated from the stream of instruction are assessments of learning. Such assessments are typically given at the end of a learning module or course as a summative assessment, to grade, rank, or categorize the learners. Formative assessments, or assessments for learning, are ones used to modify future instruction. Research shows that assessments for learning have been shown to improve learning outcomes (Black & Wiliam, 2012; Shute & Rahimi, 2017). Formative assessments help learners and their instructors gauge where the learner is compared to their learning goals, and provide appropriate supports (feedback or instruction) for them to achieve those goals.

One criticism of assessments for learning is the task of providing personalized feedback to every learner is often quite labour intensive. In a computer-supported learning environment, observations to assess the student's current state of competency can be unobtrusively captured and evaluated as the student interacts with the system. These *stealth assessments* (Shute, 2011) can be used to guide the presentation of learning support and the sequencing of learning activities. Furthermore, these types of assessments blur the boundaries of learning and assessment (Shute, 2009; Shute & Rahimi, 2017). Note that *stealth assessment* does not refer to high-stakes testing done without the learner's knowledge, but rather to low-stakes observations made with

consent in a natural problem-solving environment (often educational games or simulations) so that the system can offer formative feedback.

In a learning environment equipped with a stealth assessment, every interaction of learners with the learning environment becomes evidence of what the learner knows or can do. When a stealth assessment is used to guide feedback and learning supports, the assessment changes its negative connotation of “got you” in the eyes of the learner to become the more positive “I am here to help you.” In the end, a longitudinal collection of formative assessments in a learning environment could be used for summative purposes. Such a sequence of formative assessments could have greater reliability than a single (often time-limited) assessment given at the end of the learning unit (Almond, 2007); furthermore, because the learner loses the self-consciousness of being assessed, as they get engaged in gameplay, it may have greater validity as well.

Stealth assessment (Shute, 2011) is a top-down, automated assessment for learning which gets embedded in technology-rich learning environments (e.g., digital games) to elicit evidence for a competency of interest. As learners interact with the learning environment, their interactions get recorded in log files, which are then detected and analysed by the stealth assessment machinery and used to generate learner models (i.e., estimations of what the learners know and can do). Then, the learning environment can use those learner models to either adapt to learners' level of competency (e.g., change the level of tasks difficulty) or provide appropriate learning supports (e.g., feedback). The more this loop happens the more accurate the learner models become.

Stealth assessment uses an underlying assessment framework called evidence-centered design (ECD; Almond et al., 2015) framework of assessment. Initiated from the psychometrics field, ECD helps

with creating high-quality assessments (summative or formative). The core idea of ECD is to centre assessment design around the evidence required to make valid and reliable inferences about a learner's knowledge, skills, and abilities. When we talk about assessment quality, we usually refer to three psychometric properties: validity, reliability, and fairness (Hansen & Mislevy, 2004; Kane, 1992; Messick, 1994; Shute, 2009). Validity refers to the extent of accuracy of the estimated scores from an assessment. Reliability refers to the level of consistency of those scores. Fairness refers to the extent to which an assessment is fair for different groups of learners (e.g., genders and ethnicities). Following ECD guidelines can lead to the design of an assessment with a high level of validity, reliability, and fairness. However, the quality of these psychometric properties might change with every sample; therefore, assessment scientists should repeatedly evaluate these properties and try to improve them. In essence, ECD establishes a direct connection between claims of learner competency and the data derived from their performance. This linkage ensures the credibility and accuracy of the assessment process (Mislevy & Haertel, 2006). ECD is an iterative process with four core models: Competency Model (CM), Evidence Model (EM), Task Model (TM), and Assembly Model (AM)—see Figure 1.

The competency model (CM) serves as the foundation, clarifying the knowledge, skills, and other attributes, which are the targets of the assessment. The CM provides clear definitions of each of the variables that represent the targeted competencies, associating *claims*—predictions about what the learner can do or say—with the values of the competency variables (Almond & Mislevy, 1999). The CM also describes the distribution of competencies (particularly the correlation between competencies) in the target population and thus forms the initial version of the student model used to locate a particular learning in competency space (Almond et al., 2015; Chapter 13).

The evidence model (EM) is responsible for defining specific behaviours or *observables* that reveal the targeted competencies

within a particular *scoring context* (the learner's task within the game or simulation). In the context of a stealth assessment, we deal with two levels of observables: (1) *raw observables*: variables that are generated by the game engine as a result of learners' interactions with the game environment (shown as the circle, rectangle, and triangle shapes in the EM in Figure 1); and (2) *final observables*: variables that are accumulated into the learner models using a statistical model (shown as blue or dark rectangles in the EM in Figure 1). There are various ways to compute the final observables. For example, sometimes a raw observable is equal to a final observable. Sometimes, a summary of multiple raw observables becomes a final observable. The EM must define both: (1) how the values of the observables are set from the actions taken by the learner (*evidence rules*), and (2) how likely learners with different competency profiles are to produce specific observable values (*statistical model*; see Almond et al., 2020).

The task model (TM) outlines the characteristics of scoring contexts (assessment tasks), specifying elements that drive the difficulty (demand for skills), discrimination, and evidentiary focus (which competencies are involved). The designers aim to construct tasks that provide opportunities for the learner to exhibit the behaviours required as evidence. In the simplest case, a task is a single problem or a game level, although in more complex cases a single goal can require multiple subtasks, which each have a possibly different scoring context (Mislevy, 2024). Essentially, the TM plays a crucial role in producing assessment tasks, which can sometimes be presented as game levels, carefully crafted to align with the targeted competency variables. The TM encompasses a diverse collection of task types, forming the basis for the development of specific assessment tasks (Almond et al., 2014). Task types also can be seen as templates where task authors (e.g., game developers) can refer to and create as many instances as they want based on those templates (Rahimi et al., 2023).

The assembly model (AM) describes the final collection of tasks (e.g., game levels or test items) and how they are sequenced or

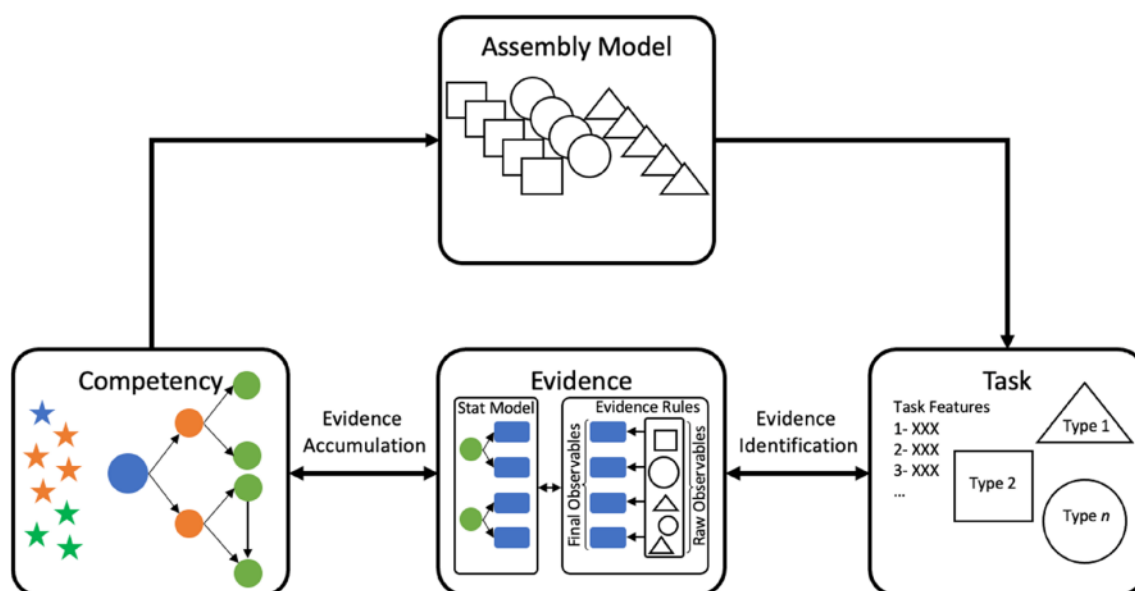


FIGURE 1 The four core evidence-centered design (ECD) models.

delivered to the learners (e.g., in an adaptive manner). The AM ensures that sufficient evidence is collected for a reliable and valid assessment. In a game-based assessment, these are often levels or challenges in the game. The reliability of a stealth assessment is determined by the number and evidentiary strength of the challenges (if we have very few observations for a competency, we can't 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/valid.

The ECD *conceptual assessment framework* (CAF) is a cohesive system where the CM, EMs, TMs, and AM work together, forming the backbone of a stealth assessment's functionality. The CAF provides the specifications for the assessment, so the design team must ensure that all the pieces (models) are in harmony. This usually requires some iteration, as design decisions made in the context of developing one model require revisiting design decisions made earlier. The CAF also contains an explicit validity argument for the assessment as it connects the collection of tasks defined in the AM, the tasks assigned through the TMs, and the evidence extracted by the EMs to the claims defined in the CM.

ECD is a theoretical framework for assessment design. Moving from theory to practice can be challenging as the same concept is often realized differently in different domains and for different assessment purposes. Each assessment design presents unique challenges and emphasizes different parts of the CAF. The first two authors of this paper teach graduate-level courses in assessment design, and in our experience working through examples is the best way to explain the connection between theory and practice (Mark Wilson, personal communication, uses the same strategy in his assessment design course).

The next section provides some approaches and considerations for a design team developing a CM. This is followed by discussions of the CM in a number of examples. The first is from a fully realized stealth assessment, the game *Physics Playground*. The remaining examples are CM developed as part of student projects in one of the assessment design classes. These examples emphasize design decisions, which were challenging for novice assessment designers. The final section discusses the challenges in CM development uncovered in the development process for these CMs and provides some guidelines for CM development. Although every stealth assessment design is unique, hopefully, the discussion of issues apparent in this set of examples will help other design teams find a way to approach the design decisions they are facing.

2 | CM DEVELOPMENT

The first step in any assessment design is identifying the target competencies and population. The assessment must be able to locate the

learner on their journey from base camp (low level on all competencies) to the top of the mountain (high level on all competencies). Just as a map does not need to mark the location of every pebble along the route, a CM should not be a comprehensive psychological examination of every competency in the domain. CM only needs to distinguish between competency profiles, which are important for future plans for the learner. Assessment designers should focus on building a model just sufficient for the *purpose* of the assessment (Almond et al., 2015). Therefore, the design process should start with establishing a clear statement of purpose.

In many assessment design projects (based on both our own experience and informal discussions with colleagues) using ECD with or without stealth assessment, the CM development process included the following four steps:

1. *Define the CM variables*: conduct an in-depth literature review or consult with domain experts to identify the competencies of interest and their facets. This defines a *competency space* of possible patterns of competency learners in the target population might have.
2. *Define the claims*: statements that predict how a learner at various places in the competency space (at certain levels of the competency variables, for example, low, medium, or high) will perform real-world tasks.
3. *Define the CM variables' population distribution*: describe both where learners in the target population of the assessment are likely to be in competency space, and especially the relationships among the competencies in the target population.
4. *Confirm the CM with the experts*: consult with domain experts who know the target population, the competency, and its pedagogy.

Although these steps are necessary, they are not necessarily done in the order presented. In our experience, as designers work through the process, they need to revisit design decisions made earlier. The CM development is iterative and continues until the competency variable definitions, the claims and the population distribution are consistent, and the experts affirm that the CM captures key variables and relationships of the competency of interest. The design issues on the other core ECD models (EM, TM, and AM) also impact the choices in the CM, so the entire ECD model development process, not just the CM development process is iterative. The following sections define the four steps of CM development in more detail.

2.1 | Define the CM variables

CM variables usually represent knowledge, skills, or abilities valued in the domain, although they can also represent other emotional or psychological constructs relevant to performance. They are almost always ordinal (sometimes with a finite number of states, sometimes continuous), but at the initial stages of CM development is often useful to discuss learners who are high, moderate, or low on a given CM variable. As the CM variables are latent, they are defined through the claims

(predictions about performance on future tasks) and operationalized by performance on the assessment tasks (e.g., in a digital game).

The final choice of CM variables and the number of states is determined by the purpose of the assessment. If the purpose of the assessment is to inform instruction (i.e., formative), different levels of the competency variable should be associated with different instructional activities or approaches. Additional variables, especially ones representing facets of another variable, may still be useful in defining the primary variable and developing assessment tasks that span the range of valued knowledge, skills, and other attributes.

The CM variable identification should be based on both a review of the literature and consultation with the experts. The domain analysis phase of assessment design described in Almond et al. (2003) includes the importance of documenting the portions of the literature and the statements of the experts that motivate the later design choices. If there are previously developed standards or learning progressions in the domain of interest, these are a good starting point as they often define instructionally relevant parts of the scale.

For example, the development process of a creativity assessment starts with a review of the literature on how creativity is usually defined. Some scholars (e.g., Acar & Runco, 2019; Runco & Acar, 2012) use *divergent thinking*, which includes *fluency*, *flexibility*, *originality*, and *elaboration*, as a proxy for assessing creativity. This is a CM for creativity with five variables, one defining overall creativity (or divergent thinking) and the other four representing facets of the high-level variable.

Another example (Almond et al., 2015; Chapter 12) has three levels. The target domain is communicative competence (in English) and represents a candidate CM developed by the assessment designers after listening to some preliminary discussions among experts in language instruction and assessment. Figure 2 presents the proposed CM variables organized into three levels. The highest level consists of overall *communication skill*. The second level corresponds to the skills identified by one of the experts: *grammar*, *correspondence competence* (using written language), *conversation competence* (using oral language), and *sociolinguistic competence* (social dimensions of language). The third level represents the four modalities of communication, *reading*, *writing*, *speaking*, and *listening*. Level 1 (overall skill) and Level 3 (four modalities) are reporting targets for the assessment, so they must be included in the model. Level 2, which represents only one view of communicative competence, may or may not be relevant to the purpose; Almond et al. (2015) also propose an alternative

model without that layer. The next logical step would be to present both proposed models to the experts to get feedback.

In the example shown in Figure 2, some of the variables on Level 3 are shared with the variables on Level 2. That is, Reading and Writing are shared between Grammar and Correspondence.

Experts are usually very comfortable with constructing hierarchical breakdowns of competencies into facets and sub-facets. It is more challenging to get them to identify facets which are common to different competencies (i.e., grammar is a part of all four modalities of communication). Often the initial expert model needs to be simplified because not enough evidence can be gathered in the available assessment window. However, the eliminated sub-facets are often useful in ensuring that the collection of assessment tasks spans the various parts of the higher-level CM variables.

2.2 | Define the claims

Claims are specific predictions of what an individual with a particular level (e.g., low, medium, or high) on one or more CM variables can do in some *future* context, either a future assessment or applying the competencies in the real world. Claims should be precise in the sense that if the learner attempts to use the competencies in the defined context and suitable observations of the performance are recorded, then it can be unambiguously determined whether or not the claim holds.

Claims are often hierarchical, with the higher-level claim having a more general context than the sub-claims. For example, the claim “Learner can write a business letter,” might include the sub-claims “Learner can write a business letter requesting information,” and “Learner can write an invoice for services rendered.” Note that the lower-level claims could be at different levels of a single CM variable, or even involve different combinations of CM variables. Due to constraints on the time available for assessment, the assessment usually must sample from the contexts represented in the claims.

2.2.1 | Organizing claims

The goal in building a CM variable is

... to build an imaginary variable along which we can define ... equal units that, although abstract in theory, are, nevertheless, approximable in practice...It is the fundamental language of science... [to] add up to reproducible meaning.

Wright, 1999, p. 67–8.

Although Wright is describing a continuous CM variable (hence the equal units), the principle remains the same if the CM variable is represented as ordinal or event binary. No matter how the CM variable is finally represented, it is helpful for designers and experts to

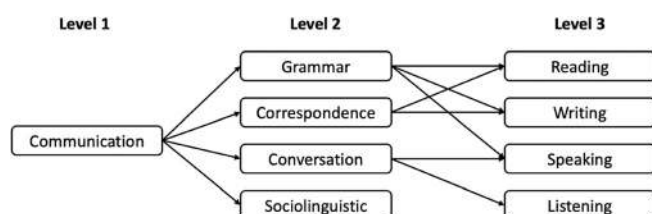


FIGURE 2 Communication competency model (CM) variables (Source: adapted from Almond et al., 2015, figure 12.5).

think about the high-end, middle, and low-end of the scale. In the CM, claims are associated with a level of the CM variable for which the claim is believed to be more likely than not to hold. At this stage, the goal is to make sure that the assessment can gather evidence to place a learner along the claims hierarchy defined by the CM variable.

A claim is *practical* if it can be assessed using a task that fits within the design constraints (testing time, cost, other resources) of the assessment. A construct map generally includes a mix of practical and impractical claims. The impractical claims are useful as they anchor the assessment designers to the kinds of real-world claims the score-users value (e.g., the learner can write a 20-page research paper). The practical claims are fodder for development into task and evidence models; in fact, as the design process progresses the claims implicit in the task and evidence models must be an extension of Almond's Rule above that there needs to be sufficient feasible claims to distinguish between levels of the competency variable.

Ultimately, the assessment must have a collection of claims (and later tasks) which will provide enough evidence to place the learner on the scale of the CM variable. There are two approaches (using a sports metaphor): *high* and *long jump*. The *high-jump* approach creates a collection of progressively more demanding tasks and infers the level of competency by the highest level at which the learner can consistently produce an adequate solution. The *long-jump* approach defines an open-ended task and infers the level of competency by the quality or quantity of the solution. Table 1 shows some claims associated with the jumping competency. Note that in a typical high-jump event, each task (attempt) is more demanding than the last as the bar is raised after each one, while in the long jump approach, all tasks (attempts) are made under the same condition. Some competencies are more naturally measured through the long-jump or high-jump approach, and some can be measured with either. Design teams should explore both approaches and choose the one that seems most natural.

As an example, consider a reading test. A high-jump approach would select a sequence of increasingly difficult (harder vocabulary, more complex sentences, less familiar topics) passages and associate the learner's reading skill with the most difficult passage they can read. A long-jump approach might provide them with a long reading passage and see how long it takes them to read it (or how much they can read in a fixed time). The long-jump approach is natural when the construct being measured includes fluency.

TABLE 1 High bar versus long jump claims.

	High bar claims	Long jump claims
High	The person can jump over 7 feet or higher bars	The person can jump to 10 feet or more
Medium	The person can jump over 5.5 feet or higher bars	The person can jump between 6 and 10 feet
Low	The person can jump over 4 feet or higher bars	The person can jump to 6 feet or less

2.2.2 | Identifying claims

In the authors' experience working with novice test designers (students in assessment design courses), identifying, and specifying claims can be difficult. Below are some strategies which may or may not be useful in any assessment design context.

Bloom's Taxonomy and Depth-of-Knowledge. Bloom's taxonomy of learning (Bloom, 1956; Krathwohl, 2002) associates various verbs with levels of a particular skill. For example, learners who can only *remember* facts or definitions but cannot apply the knowledge at hand in various situations are probably low on the CM variable; learners who can *remember* and *apply* the knowledge to different situations are probably at a medium-to-high level; and learner who can *remember* facts, *apply* the knowledge to various cases, and *evaluate* the quality of various given responses are probably at a high level.

Bloom's taxonomy uses *action verbs*, verbs which imply that information can be gathered about whether or not the task was successfully completed. Note that often the taxonomy is used for writing learning objectives, which are essentially claims. While some approaches only accept learning objectives which are observable claims, the ECD approach allows a learning objective to be a level of a latent CM variable that is defined through observable claims.

The claims chosen to be included in the assessment are further specified through the task and evidence models. Consider the claim, "The learner can successfully apply the competency in a familiar context." What is meant by "familiar context" is defined in the task model and a "successful application" is defined through the evidence model.

Consistency of performance and variety of tasks. Almost all psychometric models encode the premise that the more competency the learner possesses the more likely the learner will be to successfully complete the task. To accurately assess the learner's competency, the assessment must offer numerous opportunities for the learner to attempt to apply the competency in an appropriate context. For some competencies (particularly physical skills like throwing a ball to hit a target), repetitions of the same or essentially similar tasks might be sufficient, but in other cases, the assessment designers will need to change incidental details of the task so that the learner does not recognize the tasks as the same.

If the competency can be applied in a variety of different contexts, then learners who are higher on the scale are more likely to successfully apply the competency in a wider variety of tasks. In particular, some contexts will be more demanding than others, and for the more demanding contexts, only learners higher along the competency scale will be likely to successfully apply the competency. Generally, a learner is more likely to complete a task for which the demand is equal to or lower than that learner's competency, and less likely to complete a more demanding task. Note that tasks which have a context familiar to the learner (near transfer) are easier than tasks with an unfamiliar context (far transfer). To help span the range of the construct, the assessment designers should think about task contexts which represent near and far transfer applications of the skill.

How many claims are needed? Like many questions in psychometrics, the answer depends on the purpose of the assessment. As the

purpose of a stealth assessment is formative, the assessment must distinguish parts of the competency space corresponding to the recommendation of different instructional strategies. For those different regions (i.e., low, medium, and high levels of performance) to be different, there must be at least one claim difference between them.

The second purpose of the claims is to provide a definition of the CM variable in terms of potentially observable predictions. In this case, the number of claims is sufficient when the meaning becomes clear. One method for testing this is to have colleagues who are not part of the design team review the CM variable definitions and make sure they are clear.

Finally, note that the process of developing the conceptual assessment framework is iterative. As the design team considers the EM, TM, and AM they will naturally revisit and clarify the CM variable definitions and claims. Thus, it is possible to start with a small set of initial claims and expand the number of claims as the assessment design process unfolds.

2.3 | Define the population distribution and the CM variables' relationships

Each assignment of possible values to the CM variables defines a location in the *competency space* at which a learner could be located (e.g., low, medium, or high). Generally, instructional systems try to move learners from less desirable to more desirable parts of the competency space (e.g., from low to high). Tasks provide the most evidence about learners when the competency demands of the task are matched to the expected competency of the learners. Therefore, knowing how the learners are likely to be distributed in competency space, e.g., toward the low-end, the middle, the high end, or spread throughout, helps the assessment designers design tasks and evidence to maximize information. Certain scoring models also require a population distribution over where learners are in the competency space.

When there is more than one CM variable, then the CM must capture the relationship between (or among) the variables. In general, the variables will be correlated as learners who have spent more time studying or practicing will generally be at higher levels of the variables. In the early stages of the design process, the design team should characterize these dependencies with a knowledge graph, with nodes representing variables and edges representing relationships (see Figure 2). The edges can be labelled with the type of relationships, some common labels are described in the following paragraphs.

Part-of relationships. Often one CM variable consists of several component sub-competencies; these are *part-of* relationships. Almond et al. (2015) recommends orienting these edges from the whole to the part. For example, *Reading*, *Writing*, *Speaking*, and *Listening* are all parts of the *Communication* competency. Experts seem comfortable with hierarchical breakdowns of this kind, and in many cases the designers need to encourage the expert to think about other ways in which the CM variables could be related.

Prerequisite relationships. If learners typically master one competency before another, that is a *prerequisite* relationship. Competencies that are necessary for learning a second competency are a *hard prerequisite*, for example, letter-sound identification is a prerequisite for word-level decoding. *Soft prerequisites* are an artifact of one competency usually being required before the other, for example, addition is almost always taught before subtraction. Note that prerequisite relationships can arise in a curriculum-dependent model. For example, one world history text might put the chapter on Ancient Mesopotamian civilizations before the one on Ancient Indus Valley civilizations, and a different text might reverse the order. The relationship between knowledge about the two groups of civilizations might point in either direction depending on the chosen text. The target population of the assessment can influence the strength and direction of the relationship. For example, when children are learning their first language, usually listening and speaking are mastered first, with reading and writing coming later in childhood. For adult learners of a second language, many students learn reading and writing first as that is easier to practice without a conversation partner.

Correlational relationships. Sometimes there is a relationship between competencies which does not fall neatly into one category or the other. Two skills could be correlated because learning the first helps with learning the second and learning the second helps with the first. For example, reading vocabulary and reading comprehension skills are usually gathered at the same time with gains in one helping provide gains in the others over time. In other cases, it could just be that the two skills are taught at the same time, thus the common factor is how much instruction and practice the learner has experienced (and some learners seem to require less instruction and practice than others). Correlational edges can be oriented in either direction, although sparser graphical representations generally result when the edge is oriented from hypothesized cause to effect.

2.4 | Confirm the CM with the experts

When possible, the design team should have the CM reviewed by experts in the domain being assessed. These should be experts in both the content and the pedagogy of the domain, that is, they should understand how competency in the domain develops in the target population. For example, a researcher in Physics, who uses advanced calculus every day, may or may not understand how middle school students might approach a problem in Physics. Therefore, the expert panel should include at least some members who know the target population.

The design team should prepare alternative versions of the CM for the experts, and ask “Which one is better and why?” (Almond et al., 2017; Rahimi et al., 2023). Presenting a single model and asking “Is this version adequate?” runs up against an agreeability bias, while asking about competing models will focus the experts' attention on the differences and encourage them to verbalize their thinking. These conversations form part of the justification for the CM and should be recorded so that they can be revisited later as needed.

3 | CM DEVELOPMENT EXAMPLES

This section reviews five examples of CM development. The first is the CM for the fully realized stealth assessment in the game *Physics Playground*. The remaining four are based on class projects in a graduate course about stealth assessment taught by the first author of this paper in Spring 2022 (the students are listed as the third to the last co-authors of this paper). Each example brings up unique challenges as well as identifying competencies which novice test developers struggle to develop.

3.1 | CM development in physics playground

Physics Playground (Shute et al., 2019), a puzzle-based game, equipped with a physics engine demonstrating the laws of physics, was specifically developed to assess middle school learners' conceptual physics understanding. In *Sketching* levels (Figure 3A), learners can draw objects, create simple machines, and employ physics laws all toward reaching the goal of each level (i.e., to hit a red balloon using a green ball). In *Manipulation* levels (Figure 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.

When designing the CM of physics understanding in *Physics Playground*, we first identified standards in the Next Generation Science Standards (NGSS) relevant to the game and the target grade levels. To operationalize the physics understanding (as the competency of interest), we asked two experts to identify behaviours they expected from learners who were *high* in the competency that they did not expect from learners *medium*, or *low* in the competency (i.e., claims). Table 2 shows some claims for learners relevant to Newton's First Law using Bloom's taxonomy of learning. We did not include the claims for the entire CM of physics understanding in this paper.

These claims represent a progression of understanding and application of Newton's first law, ranging from basic recognition and comprehension at the low level to advanced analysis, prediction, and innovation at the high level of performance. 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 physics understanding. Not all relationships are hierarchical. For example, the Properties of Momentum concept is a *prerequisite* to the Conservation of

TABLE 2 Claims for Newton's first law using Bloom's taxonomy of learning method.

Level of performance	Claims
Low	1. Can recognize that objects tend to stay at rest or in uniform motion unless acted upon by an external force. 2. Can identify common situations where Newton's first law applies, such as a book remaining on a table or a moving car coming to a stop when brakes are applied. 3. Can understand that force is needed to change the state of motion and may provide simple examples of this principle.
Medium	1. Can explain in detail Newton's first law and its significance, emphasizing the concept of inertia and how it relates to everyday situations. 2. Can analyse and describe more complex scenarios where objects either maintain their state of rest or uniform motion, citing specific examples and providing clear explanations. 3. Can relate the first law to practical applications, like seat belts in cars, and discuss how the law plays a crucial role in engineering and technology.
High	1. Can predict and analyse situations where the application of Newton's first law may lead to non-intuitive outcomes, using the law to explain why objects behave as they do. 2. Can demonstrate a deep understanding of the mathematical and conceptual aspects of the first law, applying it to solve problems involving complex scenarios, including friction and varying forces. 3. Can innovate and design experiments or practical applications that rely on the principles of the first law, and critically evaluate their effectiveness and implications.

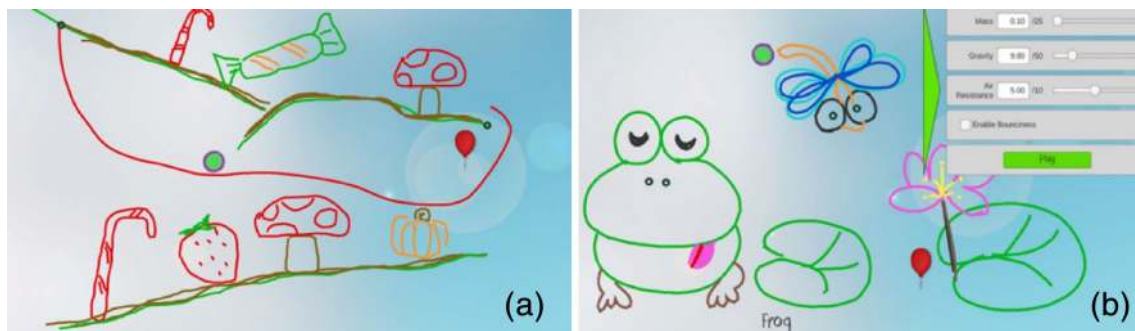


FIGURE 3 A sketching level (A) and a manipulation level (B) in *Physics Playground*.

Momentum concept 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 *correlational* relationship between these CM variables. After showing two CM versions of physics understanding to our physics experts (R. Almond et al., 2017), they selected the CM shown in Figure 4. In the following section, we discuss four other CM development examples.

3.2 | Other examples of CM development

The four CM development examples discussed in this section are selected from the *stealth assessment course* projects of the students who are included on this paper as co-authors (listed after the second author). Students formed four groups, and each designed a stealth assessment to assess four competencies: (1) systems thinking, (2) online information credibility evaluation, (3) computational thinking, and (4) collaborative creativity. In the following sections, we discuss how these groups developed their CMs and what difficulties they faced in the process.

Note that our primary aim in this paper is not to establish definitive CMs for the constructs mentioned above; instead, we acknowledge that those CMs must undergo rigorous testing for implementation. Rather, the core objective of this paper is to provide insight into the process of CM development, highlighting common challenges and offering strategies to resolve those challenges. The below examples are written by the students in each team; therefore, “we” refers to the students in each team per example.

3.2.1 | Example 1: Systems thinking

Systems thinking (ST) refers to seeing the larger picture of a system that is informed by its components, connections, and processes, but that is also more than the sum of those parts (Stave & Hopper, 2007). After conducting a literature review, we proposed a multilevel competency model with three direct sub-facets: (1) attend to elements and interconnections (Dugan et al., 2021; Stave & Hopper, 2007); (2) model the system (Shute et al., 2010); and (3) use conceptual models (Stave & Hopper, 2007). Considering that to model the system it is necessary to attend to elements and interconnections, we selected the relationship between (1) and (2) as a prerequisite relationship. At the same time, a model of the system is required to use conceptual models to describe the system. Therefore, we decided that (2) is a prerequisite to (3). Each of these three competencies has sub-facets at a third, low level (Figure 5 for the final CM for systems thinking).

To further define the CM of systems thinking, the team elaborated the possible, relevant claims for each low-level sub-facet. Table 3 includes the claims related to the sub-facet *See the Whole System*.

Difficulties and how they were resolved. The main challenges we faced when developing the ST's CM stemmed from our team's lack of expertise in ST, coupled with time constraints that prevented us from becoming ST experts. Furthermore, ST is a complex and difficult-to-define construct. It would have been advantageous to conduct a more comprehensive and in-depth systematic review before defining our model, especially in terms of gaining a deep understanding of how ST operates within the specific target population (in our case, middle school students). Alternatively, collaborating with experts in ST to design an appropriate CM would have been valuable. Unfortunately,

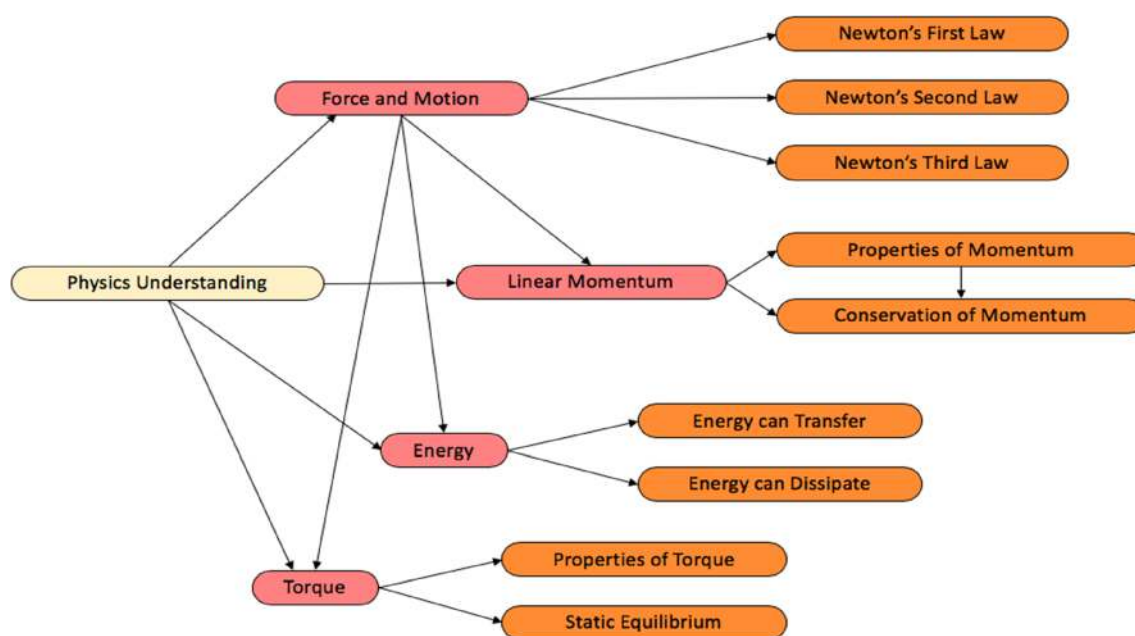


FIGURE 4 Physics understanding competency models.

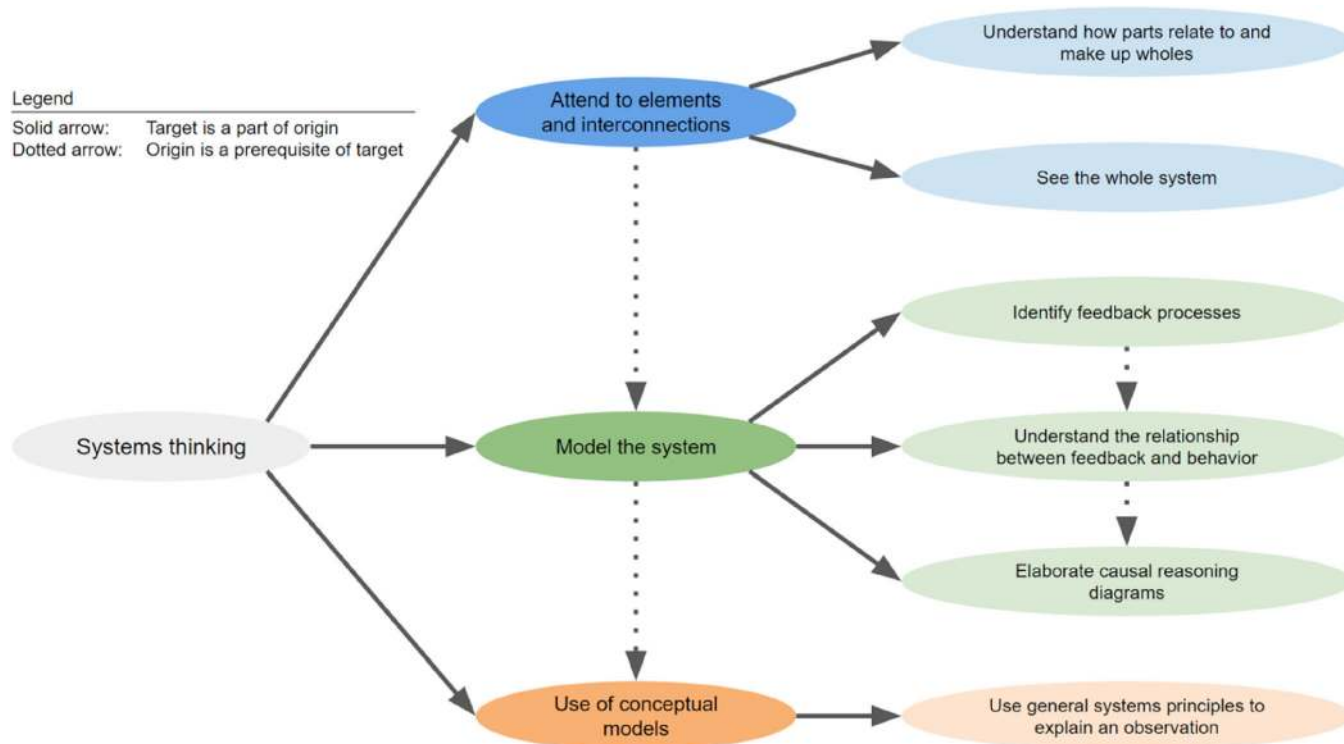


FIGURE 5 Systems thinking competency model (CM).

TABLE 3 Claims for *See the Whole System* using Bloom's taxonomy of learning objectives method.

Level of performance	Claims
Low	1. Can give an understandable name to the system according to its function.
Medium	1. Can give an understandable name to the system according to its function. 2. Can explain the system as a whole without getting stuck on details.
High	1. Can give an understandable name to the system according to its function. 2. Can explain the system as a whole without getting stuck on details. 3. Can explain the system synergy: describe properties of the system that the components alone do not explain.

due to the unavailability of such experts, we had to rely solely on existing literature.

Another challenge we encountered in developing the CM was that we created the model without a clear understanding of its ultimate application. If we had first developed a game or had more time to adjust both the game and the CM simultaneously, we might have designed the CM differently. It is essential to recognize that the design, narrative, and features of the game directly influence how we envision the CM, introducing complexity into the development process. This experience taught us a valuable lesson: in future projects,

we should consider and develop the ECD models simultaneously or iteratively.

3.2.2 | Example 2: Online information credibility evaluation

Online information credibility evaluation, which requires critically assessing and evaluating the trustworthiness of online information in various formats (e.g., websites, videos, images, and social media posts), is the most challenging aspect of information literacy (Choi, 2020; McGrew et al., 2017; Metzger et al., 2010; Sparks et al., 2016). Although targeted curricular interventions have shown the potential for positive student learning outcomes in this area (McGrew, 2020; McGrew et al., 2018), online information credibility evaluation is difficult to measure using traditional assessments because of its multidimensional competencies, which require a wide range of knowledge and skills that are difficult to transfer to authentic, real-world contexts (Shute et al. 2021; Shute & Rahimi, 2017).

To build a CM for online information credibility evaluation, we synthesized the existing literature with the help of experts that consisted of relevant assessment frameworks for measuring related information literacy competencies in secondary-student populations (e.g., Association of American Colleges and Universities, 2013; Binkley et al., 2012; Bundy, 2004; Caulfield, 2019; ETS, 2002; Environments, 2009; International Society for Technology in Education, 2022; Kempster Group, 2008; The Association of College and Research Libraries, 2015; Wineburg et al., 2016).

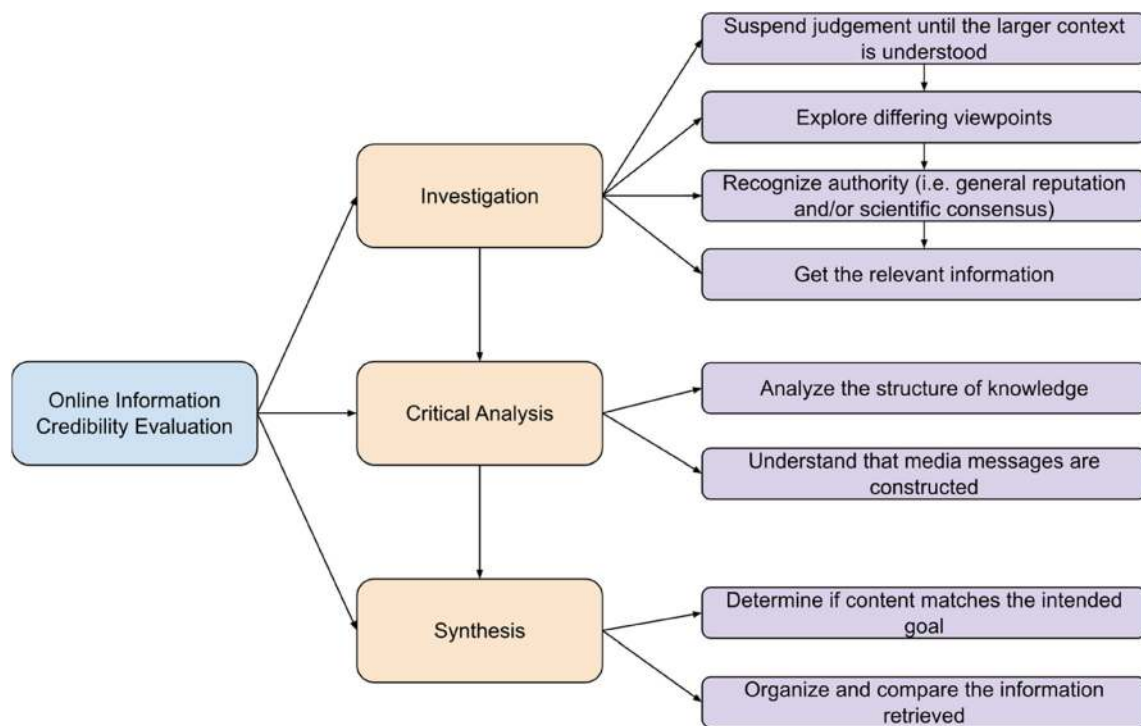


FIGURE 6 Online information credibility evaluation competency model (CM).

TABLE 4 Claims for the sub-facet *Investigation* using the consistency of performance method.

Competency	Competency level	Performance
Investigation (observables: suspending judgment until the larger context is understood; exploring differing viewpoints; recognizing authority; getting the relevant information)	Low	Rarely (0%–30% of the time) performs each indicator (e.g., reading laterally, using click restraint, etc.)
	Medium	Usually (30%–80% of the time) performs each indicator (e.g., reading laterally, using click restraint, etc.)
	High	Almost always (80%–100%) performs each indicator (e.g., reading laterally, using click restraint, etc.)

Our CM consists of three sub-facets and eight general observables. The three sub-facets are (1) investigation, (2) critical analysis, and (3) synthesis (see Figure 6 for the relationship between the sub-facets and indicators). The sub-facet *investigation* (Table 4) encompasses the process of searching for online information to yield results likely to correctly answer the evaluator's question. The sub-facet *critical analysis* encompasses the process of analysing online information for credibility. The sub-facet *synthesis* encompasses the process of compiling and organizing all of the information learned in the search,

and determining if the information found achieves the intended goal of the evaluator. The CM should be approached procedurally to obtain the most conclusive results. For instance, after investigating a source, the evaluator should critically analyse the structure and nature of the information by considering the potential for bias, decontextualization, or misrepresentation. Finally, the evaluator must synthesize the existing information to make a definitive credibility judgment. Therefore, we included the vertical, prerequisite connections between the sub-facets.

Difficulties and how they were resolved. Online information credibility evaluation is a cross-disciplinary competency with a variety of definitions and proficiencies in the broader literature including fields such as literacy, library science, educational technology, psychology, journalism, and others. Therefore, it was difficult to identify a comprehensive framework for assessing online information credibility evaluation as an independent construct dimension of information literacy. Our best option was to first generate a cohesive definition for online information credibility evaluation and then develop a multi-dimensional competency model based on a synthesis of the various frameworks, rubrics, and academic standards across disciplines and grade levels (Caulfield, 2019; Sparks et al., 2016; Wineburg et al., 2016) that we encountered in our literature review and through discussions with subject matter experts. This process was time-consuming, yet necessary, for properly identifying our competency and its relevant sub-facets, general indicators, and observables.

Another complication we had to wrestle with was that one's belief and ideology are inseparable when evaluating the credibility of information. Some evaluation frameworks tried to address this notion

by including language such as *developing and maintaining an open mind when encountering varied and sometimes conflicting perspectives, developing awareness of the importance of assessing content with a skeptical stance and doing so with a self-awareness of their own biases and world-view* (ACRL, 2015). We chose to exclude this phenomenon from our CM as it is typically conceptualized in sociocultural literacy theory literature as a mindset rather than a set of competencies that can be more easily measured.

3.2.3 | Example 3: Computational thinking

The term computational thinking (CT) was first popularized in reference to the skills postsecondary computer science students needed to be successful, defined as thought processes that are used to formulate a problem and express its solutions in such a way that a computer-human or machine can effectively carry out (Wing, 2006; Wing & Stanzone, 2016). Since then, many researchers have defined computational thinking in different ways. For instance, Aho (2012) defined it as habits of mind used to “formulate problems so students' solutions can be represented as computational steps and algorithms” (p. 832). Berland and Wilensky (2015) said it is “the ability to think with the computer-as-tool” (p. 630). Synthesizing these aforementioned definitions, we decided to demonstrate CT as the ability to formulate problems and solutions that can be effectively carried out by computers.

Based on the literature review we conducted and inspired by Kong and Wang (2021), our CM includes four sub-facets: (1) Abstracting and Modularizing: referring to the ability to conceptualize problems and divide a concept into smaller parts of code to work on (Kong & Wang, 2021); (2) Algorithms and Controls: which is the ability to understand and construct algorithms to enable problem-solving via a well-defined sequence of actions and instructions (Kong & Wang, 2021); (3) Remixing: building on and modifying the work, solutions, and ideas of others to facilitate the construction of one's own project (Kong & Wang, 2021). A student who demonstrates a high competency in remixing and reusing can incorporate the code into original programs and give attribution; and (4) Testing and Debugging: developing strategies and plans for anticipating and dealing with problems during code construction by using iterative processes to find solutions (Kong & Wang, 2021). Of these four, we believe the competencies associated with algorithms and controls are prerequisites to effectively express solutions associated with the other three sub-facets. Table 5 shows example claims for *Abstracting and Modularizing*.

Figure 7 shows the full CM with general observables per sub-facet. General observables are actions that can be observed across multiple contexts. At this stage of developing a CM, these observables can direct assessment designers to think about in-game, final observables as discussed earlier in this paper.

Difficulties and how they were resolved. Our team encountered several difficulties with the development of the CM for CT, starting with the definition of the latent trait and its sub-facets. In their CT item analysis, Kong and Wang (2021) identified a four-dimensional

TABLE 5 Claims for *Abstracting and Modularizing* using a mixture of consistency of performance and Bloom's taxonomy methods.

Level of performance	Claims
Low	1. Struggles to conceptualize and break down complex problems into smaller, manageable parts for coding. 2. Difficulty in identifying and creating modular code structures for problem-solving. 3. Rarely utilizes abstraction techniques to simplify and organize code.
Medium	1. Demonstrates an understanding of problem decomposition and can break down moderately complex problems into smaller components. 2. Able to identify and create modular code structures to address coding challenges effectively. 3. Uses abstraction techniques to simplify and organize code, enhancing clarity and maintainability.
High	1. Proficient in conceptualizing and breaking down complex problems into well-structured, smaller parts, making them amenable to coding solutions. 2. Effectively employs modular code structures to solve coding challenges, ensuring efficiency and maintainability. 3. Skillfully utilizes abstraction techniques to simplify and organize code, enhancing code readability and facilitating collaboration on projects.

conceptualization through parallel analysis. However, it remains unclear whether a student must first learn to decompose a problem (i.e., Abstracting and Modularizing) before learning how to encode a systematic series of instructions (i.e., Algorithms and Controls). Logically, students would need to first develop an understanding of algorithms as a prerequisite to modifying the codes of others (i.e., Remixing), or inspecting and improving a program (i.e., Testing and Debugging). We hypothesize that Algorithms and Controls can be understood independently of Abstracting and Modularizing. However, the Algorithms and Controls facet is a partial prerequisite of Remixing and Testing and Debugging.

We also ran into a series of issues with how we could observe the CM variables later in our stealth assessment. The block programming environment has the potential to match for measuring Algorithms and Controls, but it was more difficult to apply this toward measuring Abstracting and Modularizing. To resolve this issue, we considered adding alternative interfaces where users down recipe steps or construct flow charts. Remixing is a skill that can often be seen in collaborative online coding environments, but in our coding puzzles, there was no apparent other user with whom to collaborate. To resolve this, we plan to add multiple characters to the game with code that serves as a key to help them escape a maze or get a prize. We conjecture that the user's appropriate use, modification, and citation of code provided by non-player characters can provide evidence of remixing. In addition, the direct, implicit puzzle challenges did not immediately allow observing the sub-facet of Testing and Debugging.

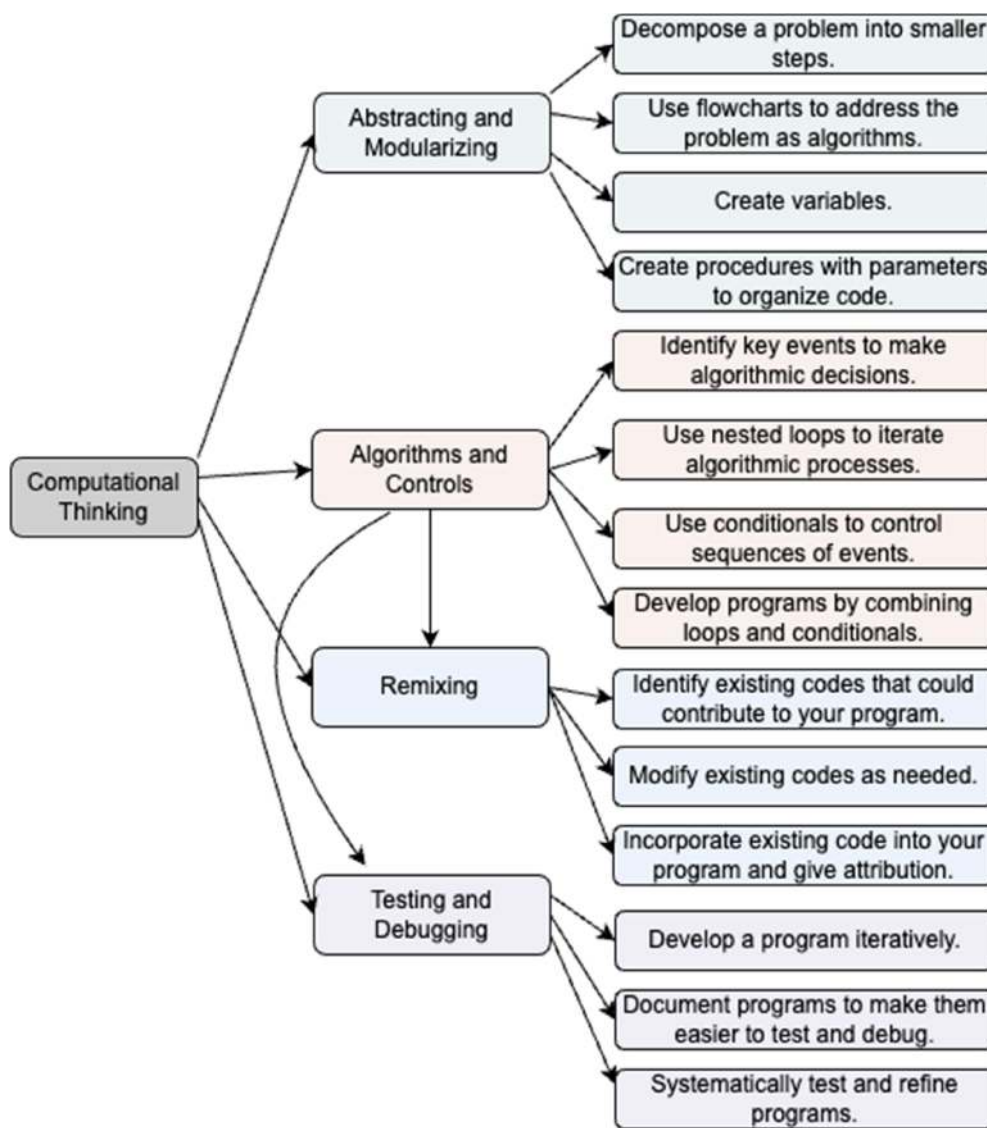


FIGURE 7 Computational thinking competency model (CM).

To resolve this, we chose to provide code with bugs purposefully placed as a challenge to allow the user to find and correct them.

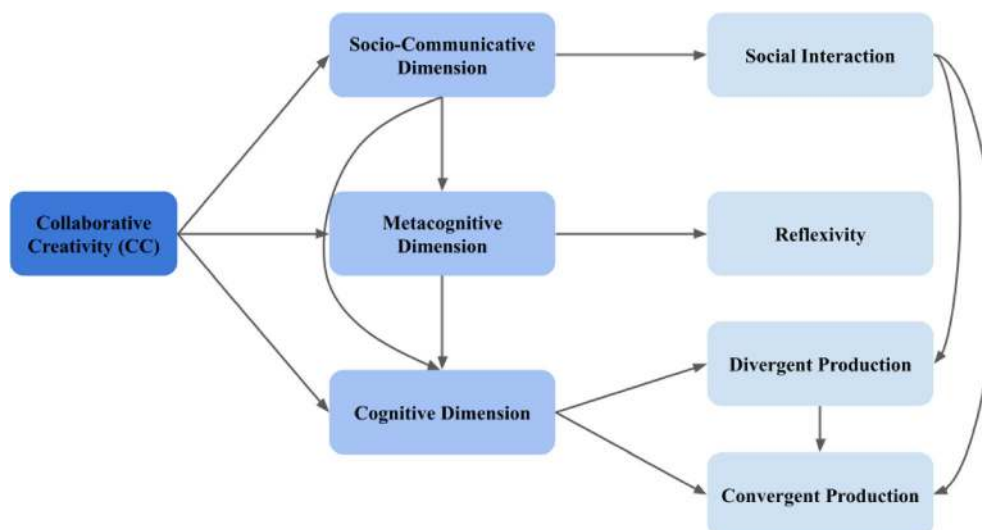
A major challenge to the incorporation of this CM into a stealth assessment is the identification of ideal responses to coding challenges. Users may take various approaches, from identifying the most efficient solution to crafting creative or purposefully complex solutions. In the latter case, the expression of user creativity might inadvertently be scored as low on the skill of Algorithms and Controls. As suggested by Buss and Gamboa (2017) and Almond (2022), a key discriminating factor could be the distinction between brute-force approaches where users attempt every possible combination of available options until one works, as opposed to finding an efficient solution in fewer steps. If an ideal solution is identified by subject-matter experts, then we can add points for similar solutions, or take away points from solutions that are further from the ideal. One possible solution is the application of natural language processing to the codes submitted by students to model ideal solutions, as

demonstrated by Min et al. (2019). However, this solution may still be limited by an abundance of efficient, but unanticipated, alternatives.

3.2.4 | Example 4: Collaborative creativity

With the challenges of sustainable development and the increased call for a culture of innovation, the development of 21st-century skills, such as creativity and collaboration, becomes essential to succeed in today's world (Thornhill-Miller et al., 2023). Interestingly, over the past two decades, several researchers have recognized that our ability to meet today's complex challenges depends on collaborative creativity (Amabile & Khair, 2008; Aragon et al., 2009; Cirella, 2016; Sawyer, 2012). Collaborative Creativity (CC) can be considered as the emergence of new, valuable, appropriate, and shared ideas between two or more people through social interactions (Elsbach & Stigliani, 2019; Glăveanu, 2018; Harvey, 2014).

FIGURE 8 Collaborative creativity competency model (CM).



Our effort to assess collaborative creativity is guided by a suite of three dimensions defined by Tan, Caleon, and colleagues (2014): (1) the socio-communicative, (2) metacognitive, and (3) cognitive dimensions. Additionally, each dimension was operationalized into one or multiple sub-dimensions (see Figure 8).

Firstly, under the socio-communicative dimension is *Social Interaction*, which describes “The group’s ability to engage in reciprocal and productive communicative interactions, including mutual grounding, that enable (rather than hinder) the preceding metacognitive and cognitive processes” (Tan, Caleon et al., 2014, p. 420). General indicators include *affectivity (non-task activity)*, *disaffectivity (non-task activity)*, and *mutual grounding (clue discussion, brainstorming)*. *Reflexivity* is a sub-dimension of the metacognitive dimension. Reflexivity refers to “The group’s ability to collaboratively self-examine, reflect on and repurpose group objectives, strategies, processes, and solutions” (Tan, Caleon et al., 2014, p. 419). The cognitive dimension consists of two sub-dimensions, *divergent production*, and *convergent production*, due to its complexity. Generally, divergent production describes the group’s ability to generate a set of ideas, methods, or solutions to address a given problem, whereas convergent production evaluates them to reach a consensus (Tan, Chai et al., 2014). Table 6 includes claims we wrote for Divergent Production.

Difficulties and how they were resolved. Attempting to design and develop a CC’s stealth assessment from scratch is challenging, especially for novices who have not learned or trained with sufficient knowledge. We as a group had run into specific issues and gone through the problem-solving process multiple times to get different models set up properly.

The first obstacle was establishing the CM after having determined our competency (i.e., CC). The team started by reviewing the literature and ended up retrieving one framework that seemed to be the most appropriate. However, there were too many sub-dimensions and related components to assemble a CM. The second issue we ran into was making connections between the evidence model (EM) and the game which was supposed to be developed to assess the competency. This issue persisted until we started thinking about the game,

TABLE 6 Claims for divergent production using the performance context method.

Level of performance	Claims
Low	Demonstrates limited ability in generating a set of ideas, methods, or solutions when addressing a problem as part of a group. May struggle to contribute to the creative process effectively, offering only a few or unvaried ideas.
Medium	Displays a moderate level of competency in generating ideas, methods, or solutions in a group setting. Can contribute a reasonable number of ideas and actively participate in creative discussions, aiding in the brainstorming process.
High	Exhibits a high level of proficiency in generating a diverse and extensive set of ideas, methods, or solutions during collaborative problem-solving. Actively contributes to creative discussions, promoting a wide range of ideas and effectively supporting the group’s creative endeavours.

but it would be more helpful if the three steps (CM, EM, and TM development) were taking place simultaneously or iteratively. We constantly revised our EM and the game design when working on TM just to make sure all of them worked cohesively to assess CC.

In general, the lack of precedent work or related literature on the design and development of stealth assessment, in specific, both models and the game design, had been the major issue for our team. We were not able to finalize the game until all three models were assembled, including the finest detail in each model (e.g., observables, associated game tasks, and evidence rules).

4 | DISCUSSION AND GUIDELINES FOR CM DEVELOPMENT

This section summarizes the challenges the design team faced and provides guidelines for designers when developing a CM.

Embrace the Complexity of CM Development. Typically, the CM development process starts with a literature review, which often takes time and effort. Note that this review must look both at how competency is defined in the domain of interest and how it is typically acquired (pedagogy). The literature typically defines the competencies in abstract terms, so building concrete claims is often challenging. Designers should anticipate that the process will require many iterations and refinements and be sure to allocate sufficient time and resources. Except for Physics Understanding, work on all the examples discussed was limited by the very short period available in a one-semester course. These examples represent intermediate drafts, and in an operational project would undergo further refinement, much of it in the context of fleshing out the rest of the conceptual assessment framework and implementing the final assessment.

Define and design all the ECD models at the “Same Time.” The complete conceptual assessment frame consists of the CM, the AM, and a collection of TMs and matching EMs; all must work together to make the complete assessment argument. There are points of contact between all of them. The AM must ensure that the collection of TMs spans all the contexts in which the constructs in the CM are measured. The TM must specify the context information for use in the AM and the work product used by the EM. The EM must connect the work product from the TM to the CM variables. Although the designers' focus may shift from one model to another, the points of contact with the other models must be included.

Papers and talks explaining ECD must always pick a sequence in which to describe each model; this can leave the impression that it is a waterfall process—one model is completed and then the next. In practice, the process is cyclical and iterative, as design decisions made in the context of one model often require revisiting earlier design decisions made in later models. Mislevy (2022), in a talk to the assessment design class, said, “We can think about the four models at the same time.” When constructing a building, the blueprint specifies how the work of the carpenters, masons, plumbers, electricians, and other builders come together to make the complete building. Similarly, the conceptual assessment framework defines the interfaces between the developers working on the game, building the tasks, and building and testing the scoring engines. When all the models are consistent, then the developers can work independently on each part of the assessment system and expect that all of the parts will work together.

The design process optimizes information about competencies within constraints: Assessment design is an engineering discipline: the goal is to optimize the amount of information the assessment provides about the learner's competencies subject to the constraints of budget, required delivery time, and the time the learner will interact with the assessment. This last constraint can have a large effect on the scope of the assessment. As a rule of thumb, there should be 3–6 pieces of evidence for each CM variable, which will be reported (additional CM variables can be useful in defining the competencies but will not be reliable enough to report. This often limits the size of the CM; it is common to prune back the initial CM).

For a stealth assessment, the setting/story and the mechanics of the game or learning environment place constraints on the types

of tasks, and hence the types of claims. In the initial version of *Physics Playground*, the mass of the ball was fixed. One of the claims was “Learner can predict how the path of the ball will change if the mass changes.” To test this claim, the game engine needed to be expanded to include a new kind of game level, where the mass of the ball can be manipulated (the domain experts considered this claim too central to Physics Understanding to support the alternative of dropping the claims from the model; Smith et al., 2023, explores this idea in more detail).

Unless the conceptual assessment framework matches the relevant parts of the domain, the assessment will not be appropriate: Identifying the right CM variables, observables to assess those variables, and tasks to produce the observables provides the first decimal place of the accuracy of the produced measure; other decisions, such as which psychometric model is used for scoring, are the second decimal place at best (Rahimi et al., 2023). There are two sets of claims associated with an assessment: the claims advertised in the score report (which may include infeasible claims far from the context of the assessment) and the claims used to develop the TM(s). The collection of tasks must provide enough evidence to establish the advertised claims with sufficient reliability and validity to meet the purpose of the assessment. No amount of statistical adjustment will fix an inadequate collection of tasks. Instead, it will be like the old joke where the man looks for his lost keys under the street light instead of out in the darkness where he lost them. Thus, getting the collection of claims right is essential to the validity of the assessment.

Consult with experts early and often. Subject matter experts, particularly experts in the pedagogy of the domain, should provide feedback on the clarity and utility of the CM variables and claims, as well as their level of agreement with the scientific consensus. The *Physics Playground* design team included two Physics instructors to provide feedback, the other four examples could be improved by expert review (not feasible within the time frame, but the next logical step). Even if experts are not available, colleagues who are not part of the design team can review for clarity. Eventually, the score reports will need to explain to learners, their instructors, and often their parents where the learner is in competency space, why the assessment made this inference, and what additional instructional steps are appropriate. A CM is complete only when it is both judged appropriate by the experts and comprehensible by the score users; checking in with the experts frequently keeps the design team on task.

5 | CONCLUSION

Although the CM is only one of the models in the conceptual assessment framework, it contains the claims which are central to the validity argument of the assessment. This paper describes the steps in developing a CM as well as some of the design considerations. This paper describes some ways of developing the claims central to the argument.

To explore how these design considerations are applied in practice, this paper describes the CM development in one fully realized stealth assessment (the game *Physics Playground*), and four initial draft

CMs built by graduate students as part of a class in assessment design. The latter examples allow some exploration of design decisions which were difficult for novice designers. Each assessment presents unique challenges, but hopefully, this collection of examples is sufficiently rich to provide other designers with possible strategies to employ in their own design.

Assessment design involves art and science in equal measure. The ECD conceptual assessment framework is the schematics for the final assessment. The CM, which contains the claims, is central to ensuring that the final assessment is suitable for its intended purpose. Stealth assessment design differs from general assessment design only in that the game mechanics and setting provide further constraints. Taking the time to design an assessment which meets the purpose and constraints can prevent wasted effort in the implementation phase.

AUTHOR CONTRIBUTIONS

Seyedahmad Rahimi: Conceptualization; methodology; data curation; software; investigation; formal analysis; supervision; visualization; project administration; resources; writing – original draft; writing – review and editing; validation. **Russell Almond:** Conceptualization; methodology; writing – review and editing; investigation; visualization; supervision; writing – original draft. **Andrea Ramírez-Salgado:** Writing – review and editing; writing – original draft; investigation; formal analysis; project administration. **Christine Wusylko:** Investigation; writing – original draft; writing – review and editing; project administration. **Lauren Weisberg:** Investigation; writing – original draft; writing – review and editing; project administration. **Yukyeong Song:** Investigation; writing – original draft; writing – review and editing; project administration. **Jie Lu:** Investigation; writing – original draft; writing – review and editing; project administration. **Ted Myers:** Investigation; writing – original draft; project administration. **Bowen Wang:** Investigation; project administration. **Xiaomaon Wang:** Investigation; project administration. **Marc Francois:** Investigation; project administration. **Jennifer Moses:** Investigation; project administration. **Eric Wright:** Investigation; project administration.

CONFLICT OF INTEREST STATEMENT

We wish to confirm that there are no known conflicts of interest associated with this manuscript.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Seyedahmad Rahimi  <https://orcid.org/0000-0001-9266-758X>

Eric Wright  <https://orcid.org/0000-0001-8963-6387>

REFERENCES

Acar, S., & Runco, M. A. (2019). Divergent thinking: New methods, recent research, and extended theory. *Psychology of Aesthetics, Creativity, and the Arts*, 13(2), 153–158. <https://doi.org/10.1037/aca0000231>

- Aho, A. A. (2012). Computation and computational thinking. *The Computer Journal*, 55(7), 832–835.
- Almond, R. (2022). Personal communication. The stealth assessment could distinguish between brute-force and more efficient approaches to problem-solving.
- Almond, R., Mislevy, R. J., & Steinberg, L. S. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–62.
- Almond, R., Shute, V., Tingir, S., & Rahimi, S. (2020). Identifying observable outcomes in game-based assessments. In *Innovative psychometric modeling and methods* (pp. 163–192). Information Age Publishing.
- Almond, R., Tingir, S., Lu, X., Sun, C., & Rahimi, S. (2017). An elicitation tool for conditional probability tables (CPT) for Physics Playground. In John-Mark Agosta and Tomas Singlair (Chair), *Bayesian Modeling Application Workshop 2017*. Symposium conducted at the meeting of Association for Uncertainty in Artificial Intelligence, Sydney, Australia. (International) Retrieved from <http://bmaw2017.azurewebsites.net/>
- Almond, R. G. (2007). An illustration of the use of Markov decision processes to represent student growth (learning). *ETS Research Report Series*, 2007(2), 1–61.
- 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. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement*, 23, 223–238.
- Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian networks in educational assessment*. Springer.
- Amabile, T. M., & Khaire, M. (2008). Creativity and the role of the leader. *Harvard Business Review*, 86, 100–109.
- Aragon, C. R., Poon, S. S., Monroy-Hernández, A., & Aragon, D. (2009). A tale of two online communities: fostering collaboration and creativity in scientists and children. In *Proceedings of the seventh ACM conference on Creativity and cognition* (pp. 9–18). ACM. <https://doi.org/10.1145/1640233.1640239>
- Association of American Colleges and Universities (AACU). (2013). *VALUE rubrics—Information literacy*. AAC&U.
- Berland, M., & Wilensky, U. (2015). Comparing virtual and physical robotics environments for supporting complex systems and computational thinking. *Journal of Science Education and Technology*, 24(5), 628–647. <https://doi.org/10.1007/s10956-015-9552-x>
- Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). Springer. https://doi.org/10.1007/978-94-007-2324-5_2
- Black, P., & Wiliam, D. (2012). Assessment for learning in the classroom. In *Assessment and learning* (pp. 11–32). SAGE Publications Ltd. <https://doi.org/10.4135/9781446250808.n2>
- Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives, handbook 1: Cognitive domain* (2nd ed.). Addison-Wesley Longman Ltd.
- Bundy, A. L. (Ed.). (2004). *Australian and New Zealand information literacy framework: Principles, standards and practice* (2nd ed.). Australian and New Zealand Institute for Information Literacy.
- Buss, A., & Gamboa, R. (2017). Teacher transformations in developing computational thinking: Gaming and robotics use in after-school settings. In P. Rich & C. Hodges (Eds.), *Emerging research, practice, and policy on computational thinking. Educational communications and technology: Issues and innovations*. Springer. https://doi.org/10.1007/978-3-319-52691-1_12
- Caulfield, M. (2019). *SIFT (the four moves)*. Hapgood.
- Choi, W. (2020). Older adults' credibility assessment of online health information: An exploratory study using an extended typology of web credibility. *Journal of the Association for Information Science and Technology*, 71(11), 1295–1307.

- Cirella, S. (2016). Organizational variables for developing collective creativity in business: A case from an Italian fashion design company. *Creativity and Innovation Management*, 25(3), 331–343. <https://doi.org/10.1111/caim.12189>
- Dugan, K. E., Mosyjowski, E. A., Daly, S. R., & Lattuca, L. R. (2021). Systems thinking assessments in engineering: A systematic literature review. *Systems Research and Behavioral Science*, 39(4), 840–866.
- Educational Testing Service. (2002). Digital transformation: A framework for ICT literacy. <https://www.zotero.org/google-docs/?broken=WgRafY>
- Elsbach, K. D., & Stigliani, I. (2019). The Physical Work Environment and Creativity: How Creative Workspaces Support and Encourage Design Thinking. In O. B. Ayoko & N. M. Ashkanasy (Eds.), *Organizational Behaviour and the Physical Environment* (pp. 13–36). Routledge: London.
- Glăveanu, V. P. (2018). Educating which creativity? *Thinking Skills and Creativity*, 27, 25–32. <https://doi.org/10.1016/j.tsc.2017.11.006>
- Hansen, E. G., & Mislevy, R. J. (2004). Toward a unified validity framework for ensuring access to assessments by individuals with disabilities and English language learners. Annual meeting of the National Council on measurement in education (NCME). <http://www.ets.org/research/dload/NCME2004-Hansen.pdf>
- Harvey, S. (2014). Creative synthesis: Exploring the process of extraordinary group creativity. *Academy of Management Review*, 39(3), 324–343. <https://doi.org/10.5465/amr.2012.0224>
- International Society for Technology in Education. (2022). *The ISTE standards*. ISTE.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527–535.
- Kempster Group. (2008). California ICT digital literacy assessments and curriculum framework. https://www.cetfund.org/wp-content/uploads/2019/09/CETF-ICT_Digital_Literacy_Assessment_and_Curriculum_Framework.pdf
- Kong, S. C., & Wang, Y. Q. (2021). Item response analysis of computational thinking practices: Test characteristics and students' learning abilities in visual programming contexts. *Computers in Human Behavior*, 122, 106836.
- Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice*, 41(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2
- McGrew, S. (2020). Learning to evaluate: An intervention in civic online reasoning. *Computers & Education*, 145, 103711. <https://doi.org/10.1016/j.compedu.2019.103711>
- McGrew, S., Breakstone, J., Ortega, T., Smith, M., & Wineburg, S. (2018). Can students evaluate online sources? Learning from assessments of civic online reasoning. *Theory & Research in Social Education*, 46(2), 165–193. <https://doi.org/10.1080/00933104.2017.1416320>
- McGrew, S., Ortega, T., Breakstone, J., & Wineburg, S. (2017). The challenge That's bigger than fake news: Civic reasoning in a social media environment. *American Educator*, 41(3), 4.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13–23. <https://doi.org/10.3102/0013189X023002013>
- Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of Communication*, 60(3), 413–439.
- Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Smith, A., Wiebe, E., Boyer, K. E., & Lester, J. C. (2019). DeepStealth: Game-based learning stealth assessment with deep neural networks. *IEEE Transactions on Learning Technologies*, 13(2), 312–325.
- Mislevy, R. J. (2024). Sociocognitive and Argumentation Perspectives on Psychometric Modeling in Educational Assessment. *Psychometrika*, 89(1), 64–83. <https://doi.org/10.1007/s11336-024-09966-5>
- Mislevy, R. J. (2022). *In a talk to the stealth assessment class*. University of Florida.
- Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered Design for Educational Testing. *Educational Measurement: Issues and Practice*, 25(4), 6–20.
- PIAAG Expert Group in Problem Solving in Technology-Rich Environments. (2009). *PIAAC problem solving in technology-rich environments: A conceptual framework (working paper)*. OECD.
- Rahimi, S., Almond, R. G., & Shute, V. J. (2023). Getting the first and second decimals right: Psychometrics of stealth assessment. In M. P. McCreery & S. K. Krach (Eds.), *Games as stealth assessments* (pp. 125–153). IGI Global. <https://doi.org/10.4018/979-8-3693-0568-3.ch006>
- Runco, M. A., & Acar, S. (2012). Divergent Thinking as an Indicator of Creative Potential. *Creativity Research Journal*, 24(1), 66–75. <https://doi.org/10.1080/10400419.2012.652929>
- Sawyer, R. K. (2012). *Explaining creativity: The science of human innovation: The science of human innovation*. Oxford University Press.
- 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). Charlotte, NC: Information Age Publishers.
- Shute, V. J., Almond, R. G., & Rahimi, S. (2019). *Physics playground (version 1.3) (computer software)*. Computer Software.
- Shute, V. J., Masduki, I., & Donmez, O. (2010). Conceptual framework for modeling, assessing and supporting competencies within game environments. *Technology, Instruction, Cognition and Learning*, 8(2), 137–161.
- Shute, V., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C., Kuba, R., Liu, Z., Yang, X., & Sun, C. (2021). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity and learning supports in educational games. *Journal of Computer Assisted Learning*, 37(1), 127–141. Portico. <https://doi.org/10.1111/jcal.12473>
- Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education: Computer-based assessment for learning. *Journal of Computer Assisted Learning*, 33(1), 1–19. <https://doi.org/10.1111/jcal.12172>
- Smith, G., Shute, V. J., Rahimi, S., Dai, C., & Kuba, R. (2023). Stealth assessment and digital learning game design. In M. P. McCreery & S. K. Krach (Eds.), *Games as stealth assessments* (pp. 81–100). IGI Global.
- Sparks, J. R., Katz, I. R., & Beile, P. M. (2016). *Assessing digital information literacy in higher education: A review of existing frameworks and assessments with recommendations for next-generation assessment* (p. 33). Educational Testing Service.
- Stave, K., & Hopper, M. (2007). What constitutes systems thinking? A proposed taxonomy. Proceedings of the 25th international conference of the system dynamics society.
- Tan, J. P.-L., Caleon, I. S., Jonathan, C. R., & Koh, E. (2014). A dialogic framework for assessing collective creativity in computer-supported collaborative problem-solving tasks. *Research and Practice in Technology Enhanced Learning*, 9(3), 411–437.
- Tan, S. C., So, H. J., & Chai, C. S. (2014). Methodological considerations for quantitative content analysis of online interactions. In B. K. Daniel (Ed.), *Handbook of research on methods and techniques for studying virtual communities: Paradigms and phenomena (volume 1)*. Information Science Reference.
- The Association of College and Research Libraries (ACRL). (2015). Framework for information literacy for higher education. <https://www.ala.org/acrl/sites/ala.org.acrl/files/content/issues/infolit/framework.pdf>
- Thornhill-Miller, B., Camarda, A., Mercier, M., Burkhardt, J.-M., Morisseau, T., Bourgeois-Bougrine, S., Vinchon, F., El Hayek, S., Augereau-Landais, M., Mourey, F., Feybesse, C., Sundquist, D., & Lubart, T. (2023). Creativity, Critical Thinking, Communication, and Collaboration: Assessment, Certification, and Promotion of 21st Century Skills for the Future of Work and Education. *Journal of Intelligence*, 11(3), 54. <https://doi.org/10.3390/jintelligence11030054>

- Wineburg, S., McGrew, S., Breakstone, J., & Ortega, T. (2016). *Evaluating information: The cornerstone of civic online reasoning*. Stanford History Education Group. <https://stacks.stanford.edu/file/druid:fv751yt5934/SHEG%20Evaluating%20Information%20Online.pdf>
- Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, 49(3), 33–35.
- Wing, J. M., & Stanzione, D. (2016). Progress in computational thinking, and expanding the HPC community. *Communications of the ACM*, 59(7), 10–11. <https://doi.org/10.1145/2933410>
- Wright, B. D. (1999). Fundamental measurement for psychology. In *The new rules of measurement* (pp. 65–104). Psychology Press.

How to cite this article: Rahimi, S., Almond, R., Ramírez-Salgado, A., Wusylko, C., Weisberg, L., Song, Y., Lu, J., Myers, T., Wang, B., Wang, X., Francois, M., Moses, J., & Wright, E. (2024). Competency model development: The backbone of successful stealth assessments. *Journal of Computer Assisted Learning*, 40(6), 2772–2789. <https://doi.org/10.1111/jcal.13025>