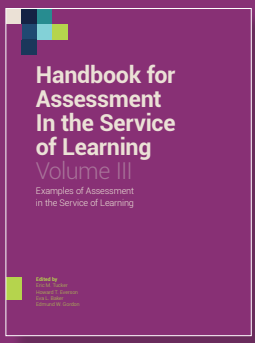
Research & Development Contributions to Assessment, Learning, Games, and Technology

Eva L. Baker and Gregory K. W. K. Chung

UMassAmherst
University Libraries



Series Editors:
Edmund W. Gordon, Stephen G. Sireci, Eleanor Armour-Thomas, Eva L. Baker, Howard T. Everson, & Eric M. Tucker

ISBN: 978-1-945764-33-2

Suggested Citation:
Baker, E. L., & Chung, G. K. W. K. (2025). Research & development contributions to assessment, learning, games, and technology. In E. M. Tucker, E. L. Baker, H. T. Everson, & E. W. Gordon (Eds.), *Handbook for assessment in the service of learning, Volume III: Examples of assessment in the service of learning.* University of Massachusetts Amherst Libraries.

Research & Development Contributions to Assessment, Learning, Games, and Technology

Eva L. Baker and Gregory K. W. K. Chung

Abstract

This chapter presents a survey of illustrative examples of CRESST's R&D contributions to assessment, learning, games, and technology. The mission of CRESST was to understand the meaning of educational quality, including approaches involving evaluation and assessment. Examples from four major areas of R&D are presented: studies of writing assessment, the assessment of rifle marksmanship, evaluation of artificial intelligence systems, and game-based learning and assessment. A foundational element of the R&D was the exploration of assessment design, development, and validation in the context of learning, both as supporting the attainment of learning goals and as an outcome measure. Every example includes the importance of designing assessments to map to the purpose of evaluation and to provide as much transparency as possible. The examples illustrate the Handbook principles of *transparency*, *purpose* and *focus*, and *validity*.

Author Note

Eva L. Baker, ORCID: https://orcid.org/0000-0001-7347-2170

Gregory K. W. K. Chung, ORCID: https://orcid.org/0000-0003-4380-5661

We have no conflicts of interest to disclose. Correspondence concerning this article should be addressed to Eva Baker, 300 Charles E. Young Drive North, SE&IS Building, Room 300, Box 951522, Los Angeles, CA 90095–1522.
Email: eva@ucla.edu

There was a time within memory when educational research and development was embraced as both important to develop new knowledge in the education and training world and for use as a scientific resource for the development of new applications intended to solve persistent problems. This chapter will highlight a few of the many contributions of the community, but it is tightly limited to a selection of work conducted at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). We describe four examples of programmatic research that took place over multiple years supported by the U.S. Departments of Education and Defense augmented by private support. The examples demonstrate CRESST's long-term commitment to designing assessments that uphold the core Assessment in the Service of Learning (AISL) principles of *transparency*, *purpose* and *focus*, and *validity*. The examples will also illustrate that developing assessment in the service of learning is not a new or abstract ideal for CRESST, but a throughline that has guided its work for decades.

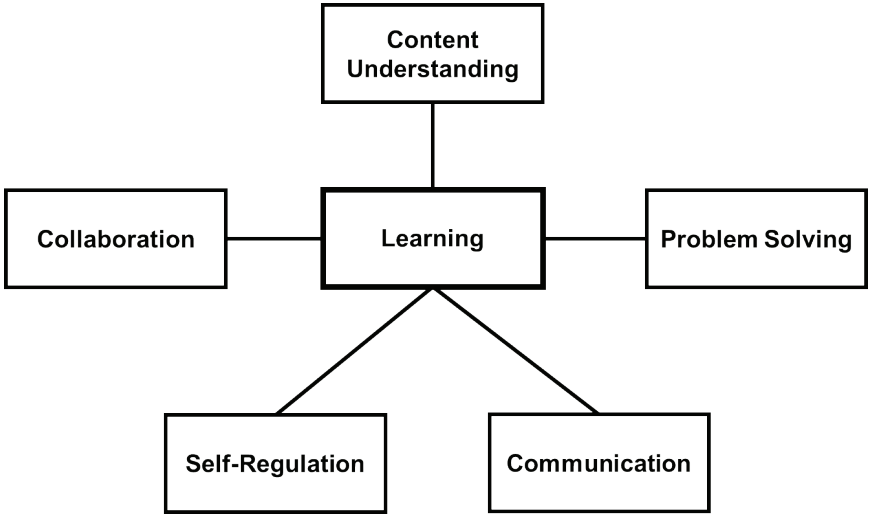
CRESST was originally developed in the mid-1960s as the Office of Education (prior to the inception of the United States Department of Education) responded to the reauthorization of Title I of the Elementary and Secondary Education Act. The response was a competition for a network of topically focused Research and Development Centers and a Network of Regional Education Laboratories focused on translation and development of usable educational options. UCLA received the 5-year award to focus on evaluation and supporting measurement and methodology in 1966 as the Center for the Study of Evaluation (CSE). Because these awards were developed to optimize the creativity of the scholars in the field, there was considerable latitude given to the design and management of research and development. When the Center grants were recompeted in 1984, CRESST was formally funded as a composite Center, where the focus was on assessment for use in schools, and partners of UCLA included universities, such as the Universities of Colorado, Illinois, and Stanford. CRESST also augmented its award with resources from state, local, federal, and private organizations.

Context of CRESST Research Design

During the period in which the research programs in this chapter occurred, three important conditions prevailed. First, the management of CRESST had extensive flexibility to select, compete, and conduct its research along with its scholars and students. It also was able to modify and adapt its objectives and procedures with little interference from the funding agencies. The ability to follow the directions of findings and to revise ongoing research plans is almost unheard of within recent funding from the federal government and as it may be in the future. Second, CRESST was a mission-focused organization. The mission of CRESST was to understand the meaning of educational quality including approaches involving evaluation and assessment. Technical studies to improve the scientific and statistical basis of the mission were an important concern, as was the exploration of alternatives to prevailing assessment approaches for policy uses. The third important element was to explore assessment in the context of learning, both as it supported the attainment of goals and as an outcome measure. In these efforts we collaborated with state and local agencies and specific organizations in the Department of Defense, including training for Army, Navy, and Marine Corps personnel.

A general model for the development of assessments was proposed and evolved over the years (Baker, 2007). Its central focus was learning supported by the various cognitive and domain requirements to promote the growth of learners. The original model, from Baker (1997), is displayed in Figure 1.

Figure 1.
Areas of Learning Identified for Model-Based Assessment (Baker, 1997)



The notion of the model also derived from research in computer science. This model was meant to be of general purpose and to be implemented in a variety of subject-matter domains. The idea of a general implementation, rather than an assessment approach that started with the subject matter, was a point of departure from traditional practice. Over the years, CRESST continued to develop and elaborate the model, for instance, using ontologies (Baker, 2007, 2012) to set boundaries for both subject matters to be included as well as the forms in which problem-solving would occur. Three criteria were developed to evaluate the quality of assessment: validity, utility, and credibility, all operating within an expectation of fairness and transparency.

The Examples

We include four examples of assessment and evaluation projects that had long-term programmatic reach. In each, we underscore the importance of learning and an understanding of both expert and learner perspectives. The principles animating this Handbook are also in play and include *transparency, purpose*

and *focus*, and *validity*. The first example we present is an effort that began with history assessment and developed into a writing assessment approach that was of general use. The second is the development of an approach to measure rifle marksmanship knowledge and skills. Both areas used expert performance as a criterion of quality as well as created models of transparent infrastructure that could be used in other assessment requirements. They were intended to focus simultaneously on learning-based assessments and outcome performance in a transparent manner. The third project focused on the development and evaluation of early versions of artificial intelligence including expert systems, natural language, and vision implementations using human benchmarking to measure the progress of AI systems. Our work also evaluated intelligent tutoring, games, and simulations. The fourth area extended R&D in learning game development and evaluation.

Simultaneously, CRESST was engaged in work in policy domains connected to local, state, federal, and international organizations focused on improving assessments, and their clarity, connection to learning and instruction, and attainment of learning goals.

Studies on Writing Assessment

This section will describe the R&D undertaken by CRESST in writing assessment. Its purpose was to apply our assessment model and develop a usable framework for the design and implementation of writing tasks to be used both in instruction and assessment of outcomes, and ultimately was generalized to other forms of constructed responses. The work involved emphases on the development of tasks to support the knowledge needed by students for writing and the ways in which scoring rubrics could be transparently designed to describe and to foster learning to write through feedback. CRESST began its interest in writing assessment in the late 1970s and focused on designs to assist state assessment agencies and to support an international study of written composition (Gorman et al., 1988). Around that time there were efforts by the Bay Area Writing Project (bawp.berkeley.edu), later the National Writing Project (www.nwp.org), to modify the way in which writing instruction took place, that is, to emphasize the process of planning, drafting, and revision. This approach also ultimately became an important part of classroom practice and assessment.

Writing Task Design: Prompt Development Supporting Prior Knowledge

We believed the writing process was only part of the solution, for our analyses and experience suggested that the design of writing tasks was not at all transparent or focused on student background. For essays to be used to evaluate content understanding, an approach was needed to capture students' prior experience. From our earlier studies, we had become convinced that students could not write well about topics on which they had little prior knowledge and that writing was not principally about appropriate style, organization, and mechanics, like punctuation and grammar, but about communicating, an approach supported by the work of Scardamalia et al. (1984). At early meetings of the IEA study on Written Composition (Gorman et al., 1988), we learned that colleagues provided content resources to writers to equalize prior knowledge and to help them flesh out their writing. CRESST staff eventually helped design tasks and scoring systems for the IEA research (Baker, 1982; Baker & Quellmalz, 1986). When CRESST was tasked by the federal government to develop secondary school history assessments, we chose to use writing as the scalable response mode to measure domain understanding. Starting with 10th grade U.S. history, we began an analysis of that content included in popular textbooks to understand student knowledge to be assessed. Unfortunately, we discovered that the treatments of important topics, such as the causes of the Civil War, were presented superficially in a paragraph of text or two and could at best provide the learner with only a thin layer of knowledge. Modeling the IEA R&D, we provided the learners with relatively short primary sources from the period of interest, using contrasting positions of politicians, for instance, the debate speeches by Abraham Lincoln and Stephen Douglas. We followed this model using opposing letters or speeches for the Revolutionary period, the Civil War, immigration in the early 20th century, and World War II among other key events in U.S. history.

Students were to read the given primary sources and then to write an essay in letter form to an absent classmate explaining the meaning of the contrasting positions. Note that over the years, we created similar assessment tasks using primary sources in history, geography, social studies, multidisciplinary topics, and science, where students read about situations and experiments rather than contrasting positions (Baker et al., 1990). In one scaled effort, we applied this approach to statewide trials in the state of Hawaii, using content in Hawaiian history and social studies topics for younger students in upper elementary school (Baker et al., 1991, 1996).

Improving on Scoring Approaches

Simultaneously, the team embarked on approaches to improve scoring by making it more transparent and valid. As noted, our interests were both outcome measures and essays assigned during courses. In both cases, the task was to improve the quality and validity of the scoring, to focus on elements that could be used for student feedback, and to reduce the time burden on teachers that scoring assigned essays imposed. The last point was critical because we had learned that teachers often severely limited the number of writing assignments given to students simply because they had no time to evaluate them. We intended to find evaluation approaches that got to the core of performance without requiring the traditional annotation and lengthy comments by teachers. Moreover, there were also approaches at the time that argued that every writing assignment required its own scoring rubric (See for example, Graves, 1978). While the idea of extracting specific information for each assignment made some sense, the reality was that teachers having to learn to use a different scoring rubric for each assignment was an incredibly unlikely outcome. Idiosyncratic scoring regimes also inhibited the ability to monitor student growth in performance over time, where a common criterion is desirable.

Do What I Do, Not What I Say

At CRESST, we decided to explore how the design of scoring rubrics could move beyond teachers' agreed-upon preferences. Our question was simple: Could we make inferences from the actual writing of experts to determine criteria for scoring student work? To that end, we asked teachers and other history experts in graduate school to write answers to prompts about epochs in U.S. history using the provided contrasting speeches. Careful analysis of the experts' writing found they organized their answers using principles or themes, they brought to bear prior knowledge external to that in the provided prompts, they used concrete examples to support their position often from the provided resources, and they avoided major mistakes or misconceptions. To use models of expertise proposed by renowned cognitive researchers (e.g., Chi et al., 1988; Ericsson & Charness, 1994; Gentner & Genter, 1983) we conducted expert-novice studies to confirm common elements in expert writing. An additional set of research involved developing and validating rater training (Quellmalz, 1982) where we focused on accuracy and speed, as we wished to support opportunities for more writing for students.

Impact and Future

The consequences of our work resulted in the development of writing approaches used for a number of state assessments, NAEP (Baker, 1981; Baker et al., 1986), and for multiyear work across literacy and mathematics domains at the elementary school level in the Los Angeles Unified School District (Niemi & Baker, 1998). We also applied these analyses to the evaluation of A level writing in Great Britain (Baker et al., 2002). Current work in AI scoring should include models generated by expert raters rather than simply interpreting identified rubrics. Our current work has focused on the identification of assessment tasks using AI-defined ontologies and domain task generation.

One of the most enduring outcomes of the studies on writing was the generality and utility of the CRESST assessment model and its emphasis on starting with learners and learning outcomes to drive the design of assessments and measures at CRESST (Baker, 1997, 2007; Baker & Gordon, 2014; Baker et al., 2022; O'Neil et al., 1990).

Assessment of Rifle Marksmanship

One of the most remarkable achievements in United States Marine Corps (USMC) marksmanship training is in developing a shooter's skill to routinely hit a 19-inch circular area at 500 yards in the prone position. The challenge posed to CRESST was to develop a way to assess marksmanship *in a distance learning context* with the goal of helping the USMC improve their non-infantry Marines' marksmanship skills.

In order to develop assessments of what was commonly believed at the time essentially a motor task, without being able to directly observe the shooter carrying out the task, required CRESST to start a program of research from first principles. Many of the methodologies developed for writing assessment were adapted for marksmanship. New frameworks and technologies needed to be developed as well, as marksmanship was never studied from an assessment perspective. In the remainder of this example, we describe the R&D program and illustrate how the domain of marksmanship was defined, how the measures were developed and validated, and how novel measurement approaches were used to explore individualizing instruction.

Determinants of Marksmanship Reexamined

At the start of the research, the marksmanship literature was focused almost exclusively on the proper execution of the motor aspects of the factors needed to establish a stable platform for the rifle and the components that underlie aiming. There was almost no conceptualization of marksmanship as a complex skill and little research to draw on to form a coherent assessment framework. To develop assessments of marksmanship that could operate under distance learning conditions, we needed to understand the underlying factors external and internal to the shooter that affected marksmanship performance.

Based on the literature and interviews with subject-matter experts (SMEs), we decomposed marksmanship performance as a function of factors within the purview of the shooter (perceptual-motor, cognitive, affective) and external to the shooter (weather, equipment). This conceptualization mirrored the CRESST assessment model (Baker, 1997, 2007) (See Figure 1). While the individual components of the model differed, how the components were identified and the role of the components as the focus for the assessments remained the same.

A key contribution was incorporating cognitive and affective components into the research. By conceptualizing marksmanship as a complex skill, we could rely on a skill acquisition model to understand how knowledge and performance interacted over time (Ackerman, 1987, 1992; Fitts & Posner, 1967). Skill development is believed to move from a learning phase to a practice phase and then to an automaticity phase. When applied to marksmanship, trainees in the learning phase are attempting to learn the concepts and rules of marksmanship. Trainees in the practice phase know what to do and practice implementing the various rules and procedures. Trainees in the automaticity phase can smoothly execute the skill with little overt consideration of the rules and procedures.

The skill model predicted the poorest performance during the learning phase when trainees are least likely to have acquired and internalized the knowledge required to shoot well (i.e., Marines who do not routinely handle weapons), suggesting measures of knowledge might be the most sensitive. For trainees in the practice and automaticity phases, perceptual-motor measures could be expected to be stronger predictors of performance. Given our population was non-infantry entry- and sustainment-level Marines, we focused on developing assessments for trainees in the learning phase and with the constraint that the assessments would need to work in a distance learning context.

Assessment Development and Validation

While we had a theoretical model of how skill develops and which phase of skill development to focus on, we needed to know precisely what knowledge Marines needed to know, how this knowledge related to shooting performance, and whether this knowledge was malleable (i.e., for applications in future distance learning training applications).

We used the CRESST assessment model to guide assessment development. We focus on identifying the cognitive demands that bear on learning, and these cognitive demands drive the design of the assessment task. The model led us to ask three questions: What are the processes (cognitive, affective, motor) that influence a trainee's successful execution of a task? What are the most direct ways of observing and measuring those processes without the measures altering the measurement itself? and How can these measures be validated to support the inferences drawn from the scores?

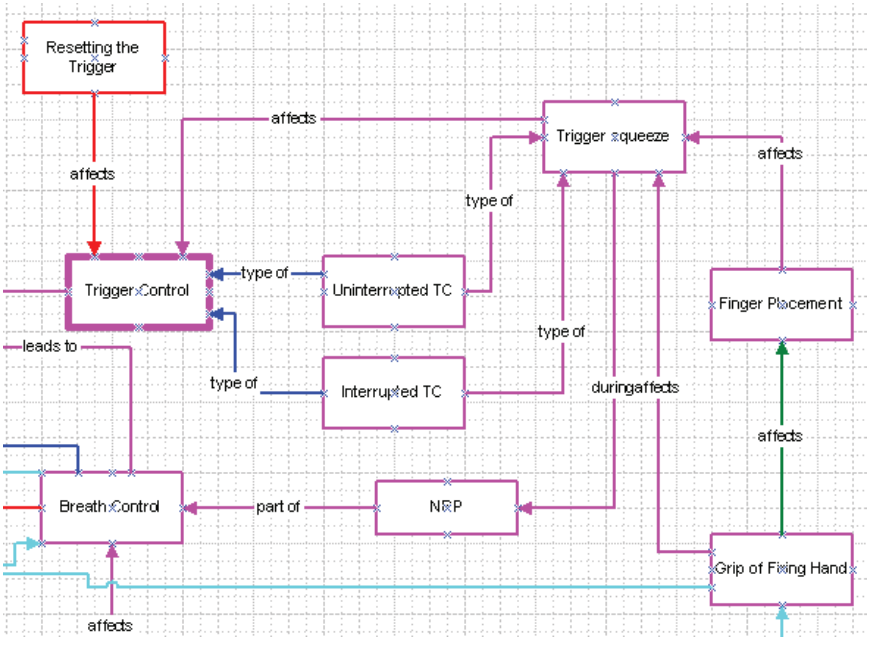
Knowledge Representations

We relied extensively on knowledge representations for practical reasons. Knowledge mapping, a method developed in the writing assessment studies to measure conceptual knowledge (Herl et al., 1996), was used to capture experts' understanding of the domain (Chung, Michiuye, et al., 2002). Experts tend to represent only the most important ideas in a domain, which is an efficient way to identify the major topic areas for an assessment. We also culled from field manuals specific cause-effect relations to augment experts' knowledge maps. The knowledge elements from experts and field manuals were stored in an ontology that was later used for scoring purposes and for instructional purposes.

Capturing Experts' Knowledge

USMC coaches and a scout sniper served as SMEs. Each SME created a knowledge map to represent how they viewed the relations among the various concepts. Figure 2 shows a fragment of the knowledge map. When we overlaid the different experts' maps, it was clear that the most sophisticated map was from a scout sniper. His map spanned multiple areas of marksmanship, reflected what we were learning from SME interviews, and presented an integrated theory of marksmanship. The differences among the various maps were consistent with USMC training, where scout snipers, compared to coaches, receive far more comprehensive and in-depth training on marksmanship.

Figure 2.
Fragment of Experts' Knowledge Maps of Rifle Marksmanship



Measures of Rifle Marksmanship Knowledge

The combination of USMC field manuals (e.g., USMC, 2001), expert interviews, and follow-up discussions with the SMEs made it clear that there was a strong knowledge component to marksmanship in addition to perceptual-motor skills. We organized this knowledge into a framework for rifle marksmanship composed of the following components: cognitive (e.g., domain knowledge), perceptual-motor (aiming, sight picture, fine and gross motor), affective (e.g., anxiety), and equipment and weather.

The set of measures we developed addressed the different components of rifle marksmanship: (a) a broad measure of marksmanship knowledge that sampled the domain and used a selected-response format; (b) a measure of conceptual knowledge using knowledge mapping; (c) an interactive task asking shooters to

identify proper and improper position elements; (d) an interactive task to interpret shot group patterns; and (e) questionnaires to survey trainees' worry, anxiety, and firing line experience. The measures went through multiple reviews by our SMEs.

Validation of Rifle Marksmanship Measures

Empirical validation tested the measures on samples with different levels of experience (non-infantry entry- or sustainment-level Marines and marksmanship coaches; high and low shooting performance) and aptitude (officer candidate school), and on trainees prior to and after instruction. In a series of three studies, we gathered evidence that, in general, suggested that the knowledge measures were sensitive to instruction, predicted record-fire scores moderately in less experienced samples, and when combined with other variables within the skill acquisition framework, predicted record-fire scores as well as scores from a rifle simulator (Chung et al., 2004). We next briefly discuss two interesting measures used in the marksmanship research: knowledge mapping and self-reported worry and anxiety.

While it was clear from the writing assessment studies that knowledge maps could be used to assess conceptual knowledge, knowledge maps were never used in a military training context. As in Herl et al. (1996), experts' maps were used as criterion maps against which trainee maps were scored. We found knowledge maps were sensitive to instruction and sensitive to expertise. Marines' knowledge map scores increased over the course of instruction (Chung et al., 2004, Study 2, 3) and Marines with more marksmanship experience scored higher than those with less experience (Chung et al., 2004, Study 2). These results are consistent with other studies that tested knowledge maps for instructional sensitivity and expert-novice differences (e.g., Herl et al., 1996, 1999; Ruiz-Primo et al., 2001).

The role of anxiety on marksmanship performance was recognized over 100 years ago. Gates (1918) reported that novice shooters' performance was affected severely by their dwelling on steadiness factors (e.g., uttering "There, I moved again"; p. 3). In our studies, the state measures of worry and anxiety administered on qualification day were among the highest predictors of record-fire score, with state anxiety and worry significantly and negatively correlating with record-fire scores (*r*s ranging from -.4 to -.5) (Chung et al., 2004, Study 2; 2005). Furthermore, when we tested the joint effects of aptitude and state worry inspired by Ackerman's (1987, 1992) study of how aptitude influences performance during the learning

phase, we found that aptitude and state worry predicted record-fire scores with a multiple *R* of .67, with state worry accounting for 34% of the variance and aptitude accounting for 11% (Chung et al., 2005).

Using Assessment to Improve Learning

Because one of our requirements was to develop assessments for a distance learning context, we anticipated the need to demonstrate how assessment information could be used for training purposes. Thus, we developed several methodologies to support future distance learning training applications given the widespread interest in the military in individualizing instruction (Bewley et al., 2009). One of the most important methodologies was the use of knowledge representations or ontologies. An ontology is domain knowledge expressed as a set of concepts and the relations that hold among the concepts (Baker, 2012; Chung et al., 2003; Gruber, 1995). Because ontologies are machine-readable and structured, software can be developed to operate on them. In our case, we created an ontology to represent marksmanship knowledge and linked instructional content in the form of text, figures, and video snippets from USMC training videos to a marksmanship concept (Chung et al., 2004). We then tested on a small sample whether individualizing instruction was effective. The results suggested that Marines receiving individualized instruction improved on topics where they initially had a knowledge gap and not on concepts they did not receive instruction on. The study strongly suggested that the methods used to model knowledge, assess knowledge, and tailor instruction were promising (Chung et al., 2003).

While we could measure one's knowledge of how to carry out a procedure (e.g., trigger control), we had no way to directly measure the execution of that skill. Our follow-on marksmanship R&D work, funded by the Defense Advanced Research Projects Agency (DARPA) investigated whether we could accelerate the acquisition of marksmanship skills. We used sensors to gather information on the difficult-to-observe processes of breath control, trigger control, and muzzle wobble (Espinosa et al., 2009; Nagashima et al., 2009) and we used an observation checklist of the various position elements considered important by experts and USMC doctrine. We tested whether we could use these fine-grained measures to (a) diagnose the novice participants' shooting problems and (b) provide effective individualized remediation using brief video-based instruction. We modeled experts' shots using the sensor data and were able to classify each

shot as expert-like or not (Nagashima et al., 2009). We found that participants who received tailored remediation significantly outperformed those who did not receive tailored instruction, with an average of 2.0 (out of 5) expert-like shots (vs. 1.0 expert-like shots). While this result may seem minor, improving novices' ability to better execute a complex skill composed of cognitive, affective, and perceptual-motor factors in 65 minutes suggested a potentially efficient approach (Chung et al., 2008).

Impact

The idea that rifle marksmanship comprises cognitive, affective, and perceptual-motor factors was novel at the time. The notion that marksmanship has a cognitive component and is a complex skill appears to be accepted by researchers worldwide as evidenced by citations to our work. The insight that marksmanship had a cognitive component was a natural development given CRESST's approach to assessment design best exemplified by Baker's (1974, 1997, 2007) focus on cognition and validity. By grounding the measurement effort around cognition and skill development, new insights were gained about which kinds of assessments would be appropriate for trainees depending on their skill development. This tailoring of measures and content was carried into instructional applications in math (e.g., Chung, Delacruz, et al., 2016), further demonstrating the utility and generality of focusing on cognitive demands first and foremost.

The second impact was the tools and methods developed or applied during the course of the research. Capturing SMEs' knowledge representation served as a method to distill the most important ideas of a domain and a way to assess learners' conceptual knowledge. The use of hardware sensors for measurement purposes would continue (e.g., Chung et al., 2021), and the conceptual and practical connection between measurement and instruction would continue to influence CRESST's technology-based R&D.

Evaluation of Artificial Intelligence (AI) Systems

AI is now at the center of attention in learning technology. We will describe a series of encounters with AI-based systems, for the most part seeking to evaluate their effectiveness. Many studies resulted in a lack of definitive findings because of the limited power of early interventions. Nonetheless, early in CRESST's history, we began numerous studies of advanced technologies, using relatively primitive implementations to explore and evaluate consequences (Baker, 1988). The story of our evaluations of artificial intelligence (AI) systems includes a few pieces. A significant note is that our work was ahead of its time; that is, it stood apart from the usual technology studies in its oddness. Only now, as AI has penetrated the daily lives of many users, our ancient studies are of renewed interest. Our evaluations included early games and simulations, expert systems and models used to support natural language processing and vision systems, and intelligent tutoring systems to promote learning. An important side effect which we will describe is our use of aspects of intelligent system design to enhance our design and implementation of assessments.

AI Games, Simulations, and Intelligent Tutoring Systems

The first game we evaluated using AI was WEST, derived from How the West Was Won, and created by Richard Burton and John Seeley Brown (Burton & Brown, 1979), titans in the early development of AI. Fascinated by the early efforts in this area, CRESST obtained support from NASA to conduct the evaluation of the game along with the Jet Propulsion Laboratory. The principal AI option in the game was a coach which was to support students' learning. We dismantled the coach, and our experiment included students who were exposed to the game with and without the coach support. The findings did not support the utility of the coach.

A second effort was supported by DARPA and was two-pronged. One set of activities was to evaluate AI-based approaches to support former service members who were afflicted with post-traumatic stress disorder (PTSD). A few private companies had created options that could be accessed through smartphones and from periods of activity and other everyday behaviors could infer episodes of PTSD and then implement support. The difficulty with this approach was that it required long periods of use as well as permissions by the users for analyses of their daily technology use. The evaluation design and beginning implementation were carried

out, but the project eventually drew no conclusions because of few users who participated for the desired length of commitment (Baker et al., 2015).

The DARPA game study ENGAGE involved the evaluation of a game developed at Carnegie Mellon University. The game was developed for primary-school-aged learners and taught children to use an adaptation of balance scales to reach conclusions about equivalence (Aleven et al., 2013). Our major evaluation finding was that games could increase the self-efficacy of young learners in the topical subject matter (Baker, 2015; Baker et al., 2016).

As part of this work, CRESST developed its own game focused on physics for 6-year-olds. The game taught concepts of mass, acceleration, and friction, where students needed to manipulate the variables to allow a train to exactly reach its station. In addition, students were to deal with bullying that occurred among characters in the game. Again, limitations of the obtained data interfered with our inferences of effectiveness. We were able to implement and further develop a framework for the evaluation of games that included cognitive demands, domain knowledge, and detailed specifications (Baker et al., 2011; Baker & Delacruz, 2016). Moreover, in developing the scenarios for the physics game, we evolved an assessment design strategy useful for creating exchangeable performance assessments efficiently. The approach created "slots" for key variables in content, task, cognitive demand, and situation that allowed the generation of comparable tasks quickly and at low cost (Baker & Delacruz, 2008).

Simulations

One outcome of our R&D around the evaluation of simulations was the development of novel measures and approaches. Simulations provide learners with experiences that might not be feasible in a classroom or training setting. The simulations CRESST evaluated required learners to engage in problem-solving and reasoning, which also meant the need for measures that would be sensitive to these higher level learning outcomes.

A persistent design goal was to measure the phenomenon in as direct a way as possible. This objective pushed R&D developments in three areas: first, to continue to apply the CRESST model of assessment, which maintained our attention on how cognitive demands of the simulation task related to the assessment task design; second, to adopt or develop measures that reflected the productive (or

nonproductive) uses of the unique learning affordances of the simulations; and third, to instrument our evaluation tools to capture and log fine-grained learner-system interactions (also called log data, trace data, or clickstream) and to use those data for assessment purposes.

Evaluating Content Understanding and Problem-Solving

Beginning in the mid-1990s, we began to explore how simulations could be used for assessment purposes. We became increasingly confident over several studies that simulations that required performance demonstrations could also be used for assessment purposes. For example, we developed a simulated web environment to evaluate middle-school students' content understanding and problem-solving. Content understanding was measured with knowledge maps, and problem-solving was measured by information seeking and search (Baker & Mayer, 1999). The educational setting was the Department of Defense Education Activity (DoDEA) middle schools in Germany, where large investments in computer-aided educational tools were introduced into the schools. The study found students' search skills and knowledge of environmental science significantly improved from the fall to spring semesters and knowledge map scores were significantly related with the quality of their search behavior (*r*s from .4 to .5) (Schacter et al., 1999).

This study was foundational in that we demonstrated the technical feasibility of collecting fine-grained behavioral process data and showed that students' online behavior was related to their content understanding and problem-solving outcomes. The capability to link students' behavior to their improved knowledge led to an obvious understanding: If students attended to the relevant content, they would learn that content. While a simplistic insight and long known in the verbal memory research, this finding was with an educationally relevant task where we could directly tie learners' behavior to the to-be-learned content. The challenge was not in the technology development or instrumentation, but rather in being able to create tasks where the learner interaction was aligned with the cognitive demands that influenced outcome performance. We concluded that under this situation, behavioral process data could be highly informative.

Given the promising results of the web search study, we then examined another simulation to gather validity evidence of the degree to which learners' online behavior reflected their cognitive processes. This linkage was important to establish because there was scant evidence in the literature to confirm that

learners' online behaviors were representations of their thinking. Establishing such a link would increase our confidence in the use of online behavior as a source of evidence about learning processes. Chung et al. (2002) collected process data and concurrent think-alouds from students as they engaged in a web-based problem-solving simulation task. The simulation required learners to determine the parents of five children (Stevens et al., 1999). The learners could access information sources with different credibility (e.g., genetic lab test results, opinions of people, library) to rule out candidate parents.

Similar to the web search results (Schacter et al., 1999), task performance was significantly and positively related to learners' fine-grained behavior reflecting the use of credible sources and negatively related to use of non-credible sources. We also confirmed that productive cognitive processes (based on students' think-alouds) were significantly related to existing validated measures of reasoning. When we examined how learners' cognitive processes were related to their online behaviors, we found that productive cognitive processing was significantly associated with task performance and productive learner behaviors and vice versa, with the magnitude of correlations in the .5 to .7 range. The results of triangulating cognitive processes derived from think-alouds, validated measures of reasoning, and learners' behaviors bolstered considerably our confidence in the use of online behavioral data for measurement purposes (Chung, de Vries, et al., 2002).

The final simulation example addressed the extent to which a simulation designed specifically for training purposes could be used for assessment purposes (Iseli et al., 2019; Savitsky, 2013). For this study, CRESST developed and validated methods to assess both declarative and procedural skills for two ultrasound-guided procedures taught in the simulator. Declarative knowledge was measured by a general test of knowledge of the two ultrasound procedures. Procedural knowledge was measured by the quality of sonographers' ultrasound scanning with a probe. The probe-motion measures were derived from moment-to-moment telemetry of the pitch, yaw, and roll of the probe. We found that more experienced sonographers demonstrated superior overall task performance and probe manipulation skills compared to less experienced sonographers, with effect sizes between the two groups of participants ranging from 0.2 to 2.0 across the various probe-based measures. These results, coupled with the marksmanship study involving sensors, suggested that the data from hardware sensers could be used in similar ways as we were using online behavior data. These results also suggested

a kind of generality: The utility of learner behavioral data is less about the specific source (software or hardware) and much more about whether the behavior is a manifestation of cognitive processes of interest.

A major theme of our simulation evaluation examples is the use of the CRESST assessment model. In every study, the learner and learning outcomes were the focus of the assessment task design effort. The cognitive demands required of the task, and in particular the unique aspects of the simulation task, guided the development of novel assessments that measured as directly as possible the presumed learning outcomes and processes. The close attention to cognitive demands and how they manifest in learners in a given task design also led to insights about which kinds of behavior in the simulation carried information related to learning and which did not. These insights would be carried into future work on game-based learning and game-based measurement.

Intelligent Tutoring Systems (ITS)

One of the most common and early uses of AI was its application to intelligent systems for learning. Early called intelligent computer assisted instruction (ICAI), several studies were conducted by CRESST (O'Neil & Baker, 1987). About two decades later these inquiries continued, supported by the Office of Naval Research (Kumar et al., 2015; VanLehn et al., 2016). In this section, instead of presenting a full example of an ITS evaluation, we present an example of measures development, a key issue when evaluating systems that individualize instruction.

The results of any evaluation rest on the quality of the outcome and process measures. ITS presents a special case because the instruction tends to be individualized, and system instructional decisions are made using granular data (e.g., presenting feedback tailored to a specific type of learner response). Thus, a challenge posed by ITSs (and systems that individualize instruction) is determining effectiveness when different students receive different degrees of content exposure, practice, and feedback.

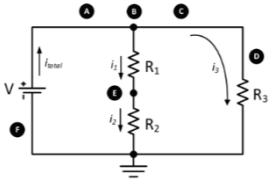
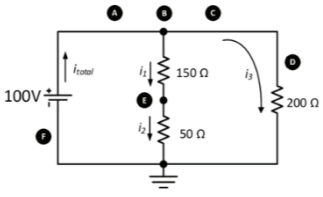
The approach we used focused on the precision of measurement. Because an ITS often attempts to remediate knowledge gaps on specific topics (e.g., understanding how to compute the equivalent resistance of three resistors in parallel), we reasoned that the measures used in evaluating the ITS should also match the precision of the instruction as a broader outcome measure might not

detect very narrow effects. One example of this approach was the evaluation of the ITS *LearnForm* (Kumar et al., 2015). *LearnForm* was an ITS problem-solving environment where students were first presented with a selected-response item. If they answered the item incorrectly, they could receive step-by-step, granular instruction and practice on the underlying topics related to the test item. The system's evaluation focused on electric circuits in AP Physics classrooms.

The measures development consisted of a physics SME first developing an ontology of electric circuits to identify the important domain concepts. These concepts were decomposed into specific knowledge components. Item development involved reviewing the electric circuit literature for misconceptions, developing canonical circuit topologies, and evaluating candidate items against the set of knowledge components.

Successful analysis of a circuit requires the simultaneous consideration of the relations among voltage, current, and resistance. To mirror this cognitive demand, we adapted an item format from Richardson et al. (1933, p. 55) and discussed in Haladyna and Rodriguez (2013). As shown in Figure 3, the item was used to assess conceptual understanding of the relations among current, voltage, and resistance, and procedural knowledge of how to apply Ohm's Law to compute voltage and current.

Figure 3.
Example Conceptual and Procedural Knowledge Items

Testing for conceptual knowledge	Testing for procedural knowledge
	
Example items	Example items [b]
1. The voltage at A is [>, =, <, indeterminate, don't know] [a] than the voltage at E.	1. The voltage at A = _____. 2. The voltage at E = _____.
2. i_1 is [>, =, <, indeterminate, don't know][a] than i_2.	3. i_1 = _____. 4. i_2 = _____.

[a] Participants select one option. [b] Participants compute the answer.

The format shown in Figure 3 allowed us to create seven scales with 41 items. The scales underwent multiple rounds of review and validation testing. The internal reliability of the scales (Cronbach's alphas) ranged from 0.7 to 0.8 (Chung, Madni, et al., 2014). Knowledge sensitivity was verified by comparing electrical engineering (EE) students to a general sample, where EE students performed significantly higher than the general sample. Instructional sensitivity of the scales was verified by first showing that the EE sample did not change over instruction (i.e., no difference in pretest and posttest scores), and also showing that scores increased from pretest to posttest in the general sample (*d*s ranging from 0.3 to 0.5). *LearnForm* effectiveness was demonstrated with an evaluation sample that improved from pretest to posttest on the scales (*d*s ranging from 0.7 to 1.9), and by demonstrating that learners who received the step-by-step instruction outperformed those who could opt out of the step-by-step instruction on the conceptual circuit analysis measure (*d* = 0.8) (Chung et al., 2015).

Human Benchmarking of AI Systems

DARPA supported an innovative set of studies evaluating early AI systems using human performance as the guide (Baker & Butler, 1991; Swigger et al., 1990). These systems included an example of natural language processing (NLP), a completed expert system in the area of scheduling, a vision system (Baker et al., 1988), and an expert systems shell. The project was initially and deliberately controversial in the computer science area, because the principal investigator was not a computer scientist. However, the evaluators of each major component came from the computer science domain. The question posed in this study was how well the system performed in comparison to human performance. Common tasks for humans were transformed and were acted upon by systems and then levels of performance were inferred. For instance, early evidence from NLP systems suggested at that time, performance was like that of a primary-school-aged child (Baker, 1994). For the most part, the work was conducted, albeit with interruptions from the funding agency when the initial supporter changed agencies. In the expert system scheduling analysis, systems managing scheduling of airplanes to gates existed, and similar tasks were given to people (O'Neil et al., 1994). Reports of this work were developed and form some of the basis of current studies of system predeveloped problem sets to evaluate comparatively the efficiency and growth of distributed systems such as ChatGPT (Baker, 1989; Baker et al., 2025).

Impact

To understand the implications of our early work in evaluating AI, two conditions are clear. One is that early formulations were extremely limited in design, and so were the evaluation options open to CRESST. To this day, CRESST is continuing to engage with AI options to support our own work in the design of ontologies and performance assessments for learning, to develop measures for various types of data collection, to explore the use of intelligent agents to act as simulated students for assessment and evaluation, and to attempt to understand what learning quality means in the era of expanding machine intelligence.

Game-Based Learning and Assessment

In this section, we present selected examples, findings, and insights from our R&D portfolio around games for learning and assessment. While the examples are drawn from our work sponsored by the U.S. Department of Education (ED), the Institute of Education Sciences (IES), and PBS KIDS, many of the methodologies and lessons learned were the result of continuous cross-fertilization among the various ongoing military games and simulation programs at CRESST sponsored by the Office of Naval Research (e.g., Baker & O'Neil, 2002; Iseli & Jha, 2016; Iseli et al., 2010; Koenig et al., 2010), DARPA (e.g., Baker et al., 2012; Baker & Delacruz, 2016; Madni et al., 2013; O'Neil et al., 2021), California Department of Education (e.g., Chung et al., 2018), private foundations (e.g., Chung, de Vries, et al., 2002), and start-up organizations (e.g., Ihlenfeldt et al., 2025).

Game-Based Learning

In 2009, CRESST was awarded a multimillion-dollar 5-year national R&D center on instructional technology grant from the U.S. Department of Education, Institute of Education Sciences (IES). The center, named the Center for Advanced Technology in Schools (CATS), developed and tested fractions math games for underperforming middle-school students in a cluster randomized controlled trial (RCT). The RCT involved 23 schools, 59 classrooms, and 1,468 students and demonstrated that students who played four fractions games performed higher on a test of fractions knowledge, compared to the comparison group who played four solving equations games (d = 0.23) (CATS, 2012; Chung et al., 2014; ED, IES, WWC, 2015). We next highlight several innovative aspects of CATS: coherent design process, game as testbed, gameplay as a data source, and advanced statistical modeling.

Coherent Design Process

We used the CRESST assessment model (Baker, 1997, 2007) to develop knowledge specifications. Ontologies were used to describe the major concepts and relations in the content domains (Baker, 2012) and the knowledge specifications succinctly described the target concepts, types of stimuli to elicit student responses, and performance expectations. The knowledge specifications standardized the requirements for assessment design, game design, and professional development for the target domains (rational number equivalence, CATS, 2013b; solving equations, CATS, 2013c; functions, CATS, 2013a). A fragment of the knowledge specification for rational number equivalence is shown in Figure 4.

Figure 4.
Snippet of Knowledge Specifications for Rational Number Equivalence

		Computational Fluency: Students can execute procedures in the domain without the need to create or derive the procedure. Fluid performance is based on recall of patterns or other well established procedures, and is fast, automatic, and error-free. *How is something done?*		**Conceptual Understanding:** Captures demonstration of understanding of the mathematical concepts. *Why is something done?*	
Rational Number Equivalence Knowledge Specifications		**When presented with… *(Assessment Stimulus)***	**Students should be able to…**	**When presented with… *(Assessment Stimulus)***	**Students should be able to…**
1.0.0. Does the student understand the importance of the unit whole or amount?					
	1.1.0. The size of a rational number is relative to how one Whole Unit is defined.	Any rational number	Place it on a number line relative to the whole interval explicitly (0 and 1 labeled) or implicitly (0 and an integer other than 1 labeled) defined.	Apparent contradictions involving rational number such as ¾ < ½ or ½ does not equal ½	Explain that the contradiction can be resolved if their relative wholes must be equal when comparing.
		A unit whole (interval, volume, area, etc.)	Show how much of the whole must be shaded to represent a fractional amount.		

All assessments, games and game levels, and professional development were designed against the knowledge specifications. Both the game levels and assessment items were mapped to the knowledge specifications, allowing verification of adequate domain coverage and alignment between the instruction, the game levels, and the assessment.

Game Testbed to Accelerate Research

A second innovation that enabled CRESST to conduct 17 design studies over two years was to design the games as a testbed. All games were designed to allow researchers to specify the level design using a text file instead of needing a programmer to program the levels. For example, in the game *Save Patch*, if a player failed the level, researchers could specify instruction or feedback tailored for the first failure, second failure, and so on, and also specify that the instruction be delivered in different modalities (e.g., text only, video). An example of the utility of the testbed was in simply modifying five text files to create five versions of *Save Patch* to identify the most promising forms of feedback to implement in the games used in the RCT (Vendlinski et al., 2011).

Gameplay as a Data Source

A third innovation was the use of fine-grained telemetry for measurement purposes. Our prior work with process data (Chung, de Vries, et al., 2002; Schacter et al., 1999) guided our telemetry design of what game mechanics to instrument, what game states to record, how to structure the data, and how to format and log the data. Yet we were unsure whether gameplay itself carried information about

learning as game-based learning was an emerging field at the time. While our first three experimental studies did not show outcome differences due to instructional variations, we did find significant gains over gameplay (*d*s from .3 to .4), hinting that the game design and game mechanics were effective in conveying the fractions concepts (Chung et al., 2010). We found that players receiving math-focused instruction (vs. game-focused instruction) generally committed fewer errors in the game that were related to math (*d*s from 0.3 to 0.5), and the math posttest was significantly related to gameplay behaviors reflecting successful fraction addition (*r*s from 0.3 to 0.6) and negatively related to gameplay behaviors reflecting unsuccessful fraction addition (*r*s around -.3). These results suggested that gameplay behavior itself carried information about learners' fractions knowledge.

These results were generally replicated in subsequent studies, suggesting that the game facilitated learners' acquisition of fractions knowledge (Vendlinski et al., 2011). Furthermore, the pattern of how gameplay related to tests of knowledge repeatedly showed that knowledge was positively related to productive gameplay behavior and negatively related to unproductive gameplay behavior, consistent with prior work (Chung & Baker, 2003; Chung, de Vries, et al., 2002; Schacter et al., 1999). These results spurred continued examination of the use of process data, including using data mining methods to detect misconceptions (e.g., Kerr, 2014; Kerr & Chung, 2012a, 2012b, 2013b), to test whether instructional variations affected specific gameplay behaviors (Buschang et al., 2012; Chung et al., 2010), to identify different learning trajectories (Kerr & Chung, 2013a), to model diagnostic assessments (Levy, 2019), and to extract best practices and guidelines on the design of telemetry (Chung, 2015). The quality of the telemetry data and RCT design, coherent game design, and external measures have led to researchers continuing to use the CATS RCT dataset to develop and explore new methods for process data analysis (Feng & Cai, 2024, 2025; Tadayon & Pottie, 2020).

Advanced Statistical Modeling

A fourth innovation was the advancement of methodology relevant to large-scale educational effectiveness studies. Cai et al. (2016) developed a novel way to account for many of the constraints inherent in multisite RCT study designs. Using the CATS RCT data, Cai et al. accounted for the RCT design constraints by using a multilevel two-tier item factor model to model latent gain. Cai et al.'s method was more precise in estimating effectiveness by being able to isolate the part of the posttest variance that was sensitive to change. The resulting effect size of $d = 0.57$

was more than twice the magnitude of the effect size computed for CATS using a classical measurement approach (d = 0.23) and used by WWC in its reviews of educational intervention studies (ED, IES, WWC, 2015).

Game-Based Assessment

The potential of using games for assessment purposes has been of interest to the measurement and assessment communities for some time (for a discussion of these issues related to games, see Baker et al., 2011; Baker & Delacruz, 2008, 2016; Delacruz, 2011; DiCerbo et al., 2016; Landers, 2015; Mislevy et al., 2015; Oranje et al., 2019; OECD, 2014, 2021; Shute & Wang, 2016; Sireci, 2016; for a discussion of these issues related to process data in assessments, see Jiao et al., 2021; Lindner & Greiff, 2023; Zumbo et al., 2023). A common aspirational goal is to "replace the dull, time-consuming, and anxiety-producing traditional approaches commonly used today" (Landers, 2015, p. vii). Landers's sentiment reflects the general desire to develop other means of measuring what learners know and can do under more engaging and complex situations.

While there may be much interest in using games for assessment purposes, numerous literature reviews have found few studies that gathered validity evidence about how games in general and game mechanics in particular relate to knowledge, skills, and learning processes (Chung & Feng, 2024; see reviews by Gómez et al., 2022; Gris & Bengtson, 2021; Kim & Ifenthaler, 2019; Tlili et al., 2021; Wiley et al., 2021). In the remainder of this section, we describe some of the R&D related to gathering such validity evidence.

Identification of Game Features That Facilitate Measurement

One of the continuous efforts in CRESST's games-related R&D has been to identify game features to support measurement. The features were identified through usability studies, qualitative feature analysis, repeated observation of similar patterns of results, and data cleaning and algorithm development. A set of the most important features are described next.

When considering a game for measurement purposes, we think the most important game feature is the alignment among the game design, game mechanics, cognitive demands evoked by the game, and the external measure used to measure the learning outcomes of the game (Baker et al., 2011; Baker & Delacruz, 2008, 2016). For example, if a game is intended to promote computational thinking, then the

gameplay should require learners to engage in the critical computational thinking processes of designing a solution, failing, debugging, and iteration. A game that minimizes learner failures and errors will not be able to detect gaps in knowledge or the presence of misconceptions because players will have few opportunities to make mistakes.

The underlying idea is that the only possible observable behaviors are the interactions the game permits. If understanding the full range of learner performance is important, then having the complement of understanding—not understanding as exhibited by errors and misconceptions—is extremely valuable because measures of success and measures of failure can provide converging validity evidence. More generally, learners with higher domain knowledge should demonstrate more productive behaviors and fewer unproductive behaviors, and learners with lower domain knowledge should demonstrate the opposite relations. We have consistently observed these complementary relations when tasks are tightly aligned with the external measures of domain knowledge (e.g., Chung & Feng, 2024).

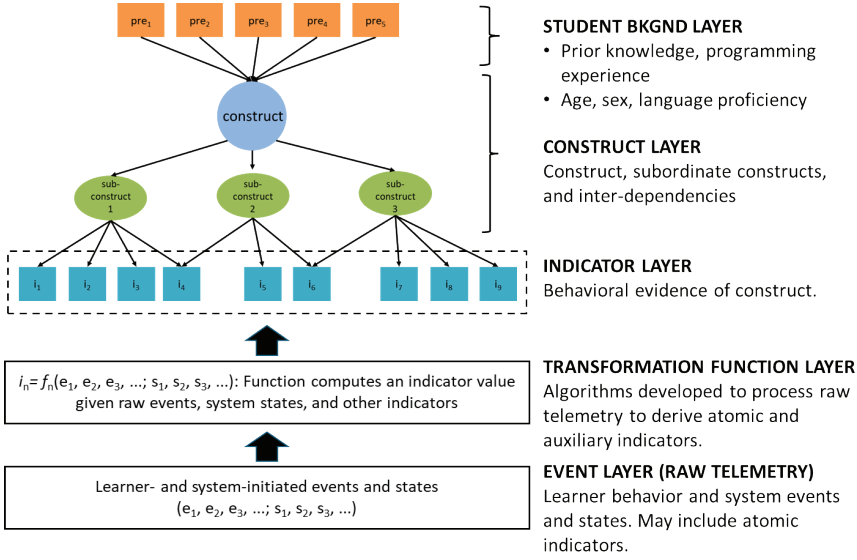
A second important game feature is practical. The user interface (UI) imposes constraints on learners' behavior (Chung & Baker, 2003). An important consideration is how to ensure that an action is intentional and not a mistake or other unwanted behavior that would contribute to construct-irrelevant variance. One type of UI element is the use of an explicit click (e.g., a button or similar UI element) that allows learners to signal, for example, that they are ready to move to the next level, to test a potential solution to a design, to select one option from a set of options, or to request help. Cleverly designed game mechanics can allow learners to perform such explicit actions as a natural part of the game. An explicit action also marks data and simplifies algorithm development by having explicit markers in the data to delineate time windows, sequences, and different levels of aggregation. Finally, game mechanics that require learners to render a judgment related to the content are especially useful if their choices can be evaluated (e.g., if moving a game piece can be evaluated as a correct or incorrect action).

Figure 5 shows how we think about fine-grained gameplay behavior (i.e., raw telemetry), indicators, and a measurement model. Indicator development often requires extensive data cleaning and processing to transform moment-to-moment events into inputs to statistical models. The programming task can range from

simply counting events to deriving numerous auxiliary variables to represent different game states that are themselves used to derive indicators. The encoding of useful information in the telemetry is dependent on both what the game allows learners to do through game mechanics, and the degree to which the game mechanics reflect the desired cognitive demands.

Figure 5.

Computational Modeling Conceptualization



Validity Evidence

Chung and Feng (2024) addressed the question, *To what extent do game-based indicators relate to criterion measures of learning*? drawing on various CRESST game-related studies. The authors reported that "common measures" composed of game performance and game progress indicators appear sensitive to the criterion measure across a broad set of games (See Chung & Feng, Appendix). The definition of game progress and game performance are game independent and analogous to the speed and accuracy variables studied extensively in verbal learning and motor learning. One use of game progress and game performance variables might also serve as a standardized metric to compare learning games

on their potential to promote knowledge or skill. See Chung and Feng (2024) and Chung and Roberts (2018) for additional examples.

The second type of indicators are game-specific indicators tailored to a game. For example, indicators of debugging behaviors were developed for a programming game (Feng & Chung, 2022), misconceptions developed for a pan balance game (Feng, 2019), deductive reasoning for a problem-solving game (Chung et al., 2018), and fractions misconceptions for a fractions game (Kerr, 2014). In all cases, the relation between the indicators and an external outcome measure were in the expected directions. Indicators that represent productive behaviors were often significantly and positively related to the external criterion measure, and indicators representing unproductive behavior were often significantly and *negatively* related to the external criterion measure (additional examples are presented in Choi, Parks, et al., 2021; Chung & Feng, 2024; Chung & Parks, 2015; Chung, Parks, et al., 2016; Redman et al., 2018, 2020, 2021, 2025; Roberts et al., 2016).

Application of Psychometric Modeling to Gameplay Data

One of the most important advances in game-based assessment was demonstrated by Feng and Cai (2024). In their study, the authors used the CATS RCT dataset to jointly model pretest, posttest, and gameplay data using a cross-classified IRT model. Feng and Cai modeled learners' latent changes in fractions knowledge and were able to directly relate the latent change to gameplay behavior. This new modeling approach directly provides information often of most interest in educational interventions: How much did learners learn (as described by latent changes in learners' knowledge over the course of instruction), and what variables influenced their learning (as described by learners' gameplay behavior)? Furthermore, the modeling technique is sufficiently general to incorporate other streams of data, such as multimodal data (e.g., eye tracking), learner background information, level design information, and interactions between learners' characteristics and the instructional setting.

Use of Population Data

One challenge presented by PBS KIDS (See Roberts et al., in press) was to examine how games played "in the wild" (i.e., the population) can be used to understand PBS KIDS' audience better. The only information available with population gameplay data is an anonymous ID. Three general issues were explored: using

psychometric modeling to estimate latent ability, using population-derived models and parameters in RCT studies, and testing a method to infer learning solely from players' gameplay behavior.

Psychometric Modeling of Population Gameplay Data

In numerous studies involving PBS KIDS' gameplay data from players "in the wild," CRESST applied various psychometric models. A close analysis of the game design and available gameplay indicators dictated the choice of models. The models included higher order IRT (de la Torre & Song, 2009) and diagnostic classification (Rupp et al., 2010) in Choi, Suh, et al. (2021); Rasch and Rasch Poisson counts (Rasch, 1960), IRT trees, and linear logistic testing model (De Boeck & Wilson, 2004) in Redman et al. (2021); a one-factor 2PL model, bifactor 2PL model with two and three specific factors in Redman et al. (2023); a multiple-group two-time point nominal IRT model (Cai, 2010; Cai & Houts, 2021) in Redman et al. (2025); and a two-time point graded response IRT model in Feng et al. (2025).

Using Population Information in RCT Studies.

To demonstrate how population data could be used in RCT studies, Choi, Parks, et al. (2021) used population gameplay to fit higher order IRT models for two PBS KIDS games. Choi, Suh, et al. (2021) used the population-based models and estimated model parameters from Choi, Parks, et al. (2021) to estimate ability of learners playing the same games in an RCT sample (Education Development Center, Inc., & SRI International, 2021). Diagnostic classification models (DCM) were also used to estimate informational text attribute profiles in both the RCT sample and population.

Estimating Learning in the Population Through Gameplay.

Finally, we explored the use of PBS KIDS games played "in the wild" to directly measure changes in gameplay that were consistent with changes in learning (Redman et al., 2023). The games were classified into three categories (likely, less likely, not likely) on their potential to promote learning. A two-timepoint latent variable model was used to estimate changes in latent ability using only game-based indicators. The study found that for the two games rated as not likely or less likely to result in learning, the effect sizes of the change in latent score were 0.07. In contrast, for the two games that were rated as likely to result in learning, the effect sizes of the change in latent score were 0.56 and 0.59.

Impact

The breadth of CRESST R&D around games for learning and games for assessment have led to insights about the conditions needed for both learning and measurement to be realized: Games that are effective in promoting learning can also yield information about learners' knowledge and skills, but only if (a) the game design and game mechanics in particular evoke the intended cognitive demands, (b) the game is instrumented to collect moment-to-moment telemetry and game state information, (c) the algorithms used to derive indicators from the telemetry are able to represent a range of performance, and (d) the psychometric models account for the constraints imposed by the game itself.

An important implication of this work for AISL is the idea of *measurement without testing*. Regardless of the type of task—game or otherwise—if the learner's behavior in the task is a manifestation of the desired cognitive demand, then the learner's behavior can serve as evidence of the cognitive demand occurring. This idea holds regardless of whether a task is designed for testing purposes or for learning purposes, for it is the interaction that is the atomic unit of observation.

Conclusion

This chapter presented a few examples of CRESST research extending over several years of effort and gave only a handful of references for each of them. Every area includes the importance of designing assessments to map to the purpose of evaluation and to provide as much transparency as possible. In most cases, our evaluations addressed not only performance on outcomes, but the value of instructional procedures and learner processes as well.

CRESST did not always juggle well the competing goals of innovation and early involvement with longer term impact. Much of our work was, in a self-aggrandizing sense, ahead of its time. This lack of fit with the context of learning and assessment vastly limited its immediate impact. However, we want to acknowledge and thank those educational and technology leaders who joined with us to explore learning and assessment strategies that were often too early for widespread use. There are numerous examples of other CRESST activities that affected proximal practice. The selection we chose to highlight, however, are focused on ideas that continue to affect educational research and development.

The methodologies and insights described in the examples also foreshadow the movement toward AISL, most clearly seen in the focus, since the inception of CRESST, on exploring assessment in the context of learning to support both attainment of learning goals and as an outcome measure. As the examples illustrate, designing assessments in the context of learning:

- Emphasizes measuring the most important concepts and skills.
- Conceives of human performance as being on a continuum, which naturally leads to the choice of experts as the criterion or reference against which to judge learner performance.
- Situates cognitive demands as a core assessment design requirement. By specifying and unpacking the key learning processes and outcomes a task is expected to evoke from learners, the assessment design process can be focused. Clear specifications can guide the development of measures, instructional content, and professional development.
- Treats the quality of measures as a necessary condition for drawing valid inferences by having clear and comprehensive definitions of what is to be measured, by making explicit how a student response is transformed into a quantitative value, and gathering validity evidence that the measures behave in expected ways.
- Is agnostic on the instructional or assessment setting, as well as the media, mode, and format used for instruction or assessment. Paper, digital, selected-response or constructed-response modes and formats can provide different information under different situations.
- Does not preclude a learning task from providing measurement information. A learning task can provide information about learners' ongoing knowledge and skills if learners are able to actually engage in the target cognitive demands and if learners' behaviors can be captured and stored.

As the assessment enterprise moves increasingly toward AISL, we think CRESST's experience can shed light on some of the challenges and opportunities ahead. The most important challenge is an understanding of cognitive demands and its implications for task design, the types and range of learner responses evoked by the task, and data capture opportunities. Additionally, adopting a naive view of measurement may be helpful for alignment, especially in technology-based

environments. If we think of the initial stages of measurement as simply an observation with some quantitative value assigned to it, then we can view a task as a set of learner-system interactions. Most of the interactions will be of little interest, but interactions that reflect judgment, decision making, or application of the target knowledge can be highly informative because they presumably reflect the outputs of learners' knowledge and skill. Furthermore, these interactions can be thought of as atomic units that can be combined, sequenced, or aggregated to form indicators that match future claims and inferences. Finally, this conceptualization, used in our work in simulations and games, can be applied to any environment where interactions exist. The limiting factor is observational capability.

The examples in this chapter addressed the Handbook principles of *transparency*, *purpose* and *focus*, and *validity.* As the field moves to more technology-based solutions, we think these principles become even more salient. Complex technology often obfuscates what is actually happening "under the hood" making independent inspection and critique nearly impossible. One path to make such systems more transparent is to develop tools and methods to specify in a formal way what to measure and the rules for transforming an observation into a measure. Another path is the training of assessment designers and technology developers on the AISL principles, methodologies, and insights described in this chapter so that best practices are designed into the applications. Regardless of approach, we are confident that AISL can be realized.

References

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin, 102,* 3–27.

Ackerman, P. L. (1992). Predicting individual differences in complex skill acquisition: Dynamics of ability determinants. *Journal of Applied Psychology, 77,* 598–614.

Aleven, V., Dow, S., Christel, M., Stevens, S., Rosé, C., Koedinger, K., Myers, B., Flynn, J. B., Hintzman, Z., Harpstead, E., Hwang, S., Lomas, D., Reid, C., Yannier, N., Fathollahpour, M., Glenn, A., Sewall, J., Balash, J., Bastida, N., & Zhang, X. (2013). Supporting social-emotional development in collaborative inquiry games for K-3 science learning. In *Proceedings of the 9th Games+ Learning+ Society Conference-GLS* (Vol. 9, pp. 53–60). ETC Press.

Baker, E. L. (1974). Beyond objectives: Domain-referenced tests for evaluation and instructional improvement. *Educational Technology, 14*(6), 10–16.

Baker, E. L. (1981, April 13–17). *Issues in the evaluation of composition instruction* [Paper presentation]. Annual meeting of the American Educational Research Association, Los Angeles, CA, United States.

Baker, E. L. (1982). The specification of writing tasks. *Evaluation in Education: An International Review Series*, *5*(3), 291–297.

Baker, E. L. (1988). Evaluating new technology: Formative evaluation of intelligent computer assisted instruction. In R. J. Seidel & P. D. Weddle (Eds.), *Computer-based instruction in military environments* (pp. 155–162). Plenum Publishing.

Baker, E. L. (1989, March 27–31). *The role of outcome measurement in the development and assessment of AI-based educational systems* [Paper presentation]. Annual meeting of the American Educational Research Association, San Francisco, CA, United States.

Baker, E. L. (1994). Human benchmarking of natural language systems. In H. F. O'Neil & E. L. Baker (Eds.), *Technology assessment of software applications* (pp. 85–98). Erlbaum.

Baker, E. L. (1997). Model-based performance assessment. *Theory Into Practice, 36*(4), 247–254.

Baker, E. L. (2007). Model-based assessments to support learning and accountability: The evolution of CRESST's research on multiple-purpose measures. *Educational Assessment* (Special Issue), *12*(3&4), 179–194.

Baker, E. L. (2012). *Ontology-based educational design: Seeing is believing* (Resource Paper No. 13). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L. (2015). *Final report: Gamechanger: Using Technology to Improve Young Children's STEM Learning* (Deliverable to funder). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., Baron, J., Coffman, W., Freedman, S., Quellmalz, E., & Williams, P. (1986, September). *Issues in writing assessment for NAEP* [Paper presentation]. Office of Technology Assessment, Washington, DC.

Baker, E. L., & Butler, F. A. (1991). *Artificial intelligence measurement system: Overview and lessons learned* (ED332677). ERIC. https://files.eric.ed.gov/fulltext/ED332677.pdf

Baker, E. L., Choi, K., & O'Neil, H. F. (2022). The training assessment framework: Innovative tools using scenario-based assessment and feature analysis. In A. M. Sinatra, A. C. Graesser, X. Hu, B. Goldberg, A. J. Hampton, & J. H. Johnston (Eds.), *Design recommendations for intelligent tutoring systems* (pp. 31–39). U.S. Army Combat Capabilities Development Command.

Baker, E. L., Chung, G. K. W. K., & Delacruz, G. C. (2011). The best and future uses of assessment in games. In *Technology-based assessments for 21st century skills* (pp. 227–246). Information Age Publishing.

Baker, E. L., Chung, G. K. W. K., Delacruz, G. C., & Griffin, N. C. (2012). *Engage validation plan* (Year 1 Deliverable). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., Clayton, S., Aschbacher, P., Chang, S., & Ni, Y. (1990, April 16–20). *Measuring deep understanding of history: The integration of prior knowledge and knowledge acquisition in explanations* [Paper presentation]. Annual meeting of the American Educational Research Association, Boston, MA, United States.

Baker, E. L., & Delacruz, G. C. (2008). A framework for the assessment of learning games. In H. F. O'Neil & R. S. Perez (Eds.), *Computer games and team and individual learning* (pp. 21–37). Elsevier.

Baker, E. L., & Delacruz, G. (2016). A framework to create effective learning games and simulations. In H. F. O'Neil, R. S. Perez, & E. L. Baker (Eds.), *Using games and simulations for teaching and assessment: Key issues* (pp. 3–20). Routledge/ Taylor & Francis.

Baker, E. L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 131–153). Prentice-Hall.

Baker, E. L., & Gordon, E. W. (2014). From the assessment OF education to assessment FOR education: Policy and futures. *Teachers College Record*, 116(11), 1–24.

Baker, E. L., Koenig, A. D., Lee, J. J., Choi, K., O'Neil, H. F., Michiuye, J., K., & Griffin, N. (2025, January 4–7). *The Training Assessment Framework project: Flexible design from classroom to AI* [Paper presentation]. Annual Hawaii International Conference on Education, Honolulu, HI, United States.

Baker, E. L., Lee, J. J., Rivera, N. M., Choi, K., Bewley, W. L., Stripling, R., O'Neil, H. F. Jr., & Redman, E. (2015). *Detection and computational analysis of psychological signals: Evaluation of digital library and SimSensei for veteran use* (Final deliverable to funder). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., Lindheim, E. L., & Skrzypek, J. (1988). *Directly comparing computer and human performance in language understanding and visual reasoning* (CSE Report 288). University of California, Los Angeles, Center for the Study of Evaluation.

Baker, E. L., Madni, A., Choi, K., Kim, J., Redman, E. H., Delacruz, G. C., Chung, G. K. W. K., Griffin, N. C., & O'Neil, H. F. (2016). *ENGAGE2: Computer games for science learning at early grades: Evaluation, assessment, and neuro-sensing investigation* (Deliverable to funder). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., & Mayer, R. E. (1999). Computer-based assessment of problem solving. *Computers in Human Behavior, 15*(3–4), 269–282.

Baker, E. L., McGaw, B., & Sutherland, S. (2002). *Maintaining GCE A level standards*. Qualifications and Curriculum Authority.

Baker, E. L., Niemi, D., Herl, H., Aguirre-Muñoz, Z., Staley, L., Linn, R. L., & Rogosa, D. (1996). *Report on the content area performance assessments (CAPA): A collaboration among the Hawaii Department of Education, the Center for Research on Evaluation, Standards, and Student Testing (CRESST) and the teachers and children of Hawaii* (Final Deliverable). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., & O'Neil, H. F. (2002). Measuring problem solving in computer environments: Current and future states. *Computers in Human Behavior*, *18*(6), 609–622. https://doi.org/10.1016/S0747-5632(02)00019-5

Baker, E. L., & Quellmalz, E. (1986, March 13–15). *Initial results for the U.S. writing study* [Paper presentation]. Conference on College Communication and Composition, New Orleans, LA, United States.

Bewley, W. L., Chung, G. K. W. K., Delacruz, G. C., & Baker, E. L. (2009). Assessment models and tools for virtual environment training. In D. Schmorrow, J. Cohn, & D. Nicholson (Eds.), *The PSI handbook of virtual environments for training and education: Developments for the military and beyond* (pp. 300–313). Praeger Security International.

Burton, R. R., & Brown, J. S. (1979). An investigation of computer coaching for informal learning activities. *International Journal of Man-Machine Studies*, *11*(1), 5–24. https://doi.org/10.1016/S0020-7373(79)80003-6

Buschang, R. E., Kerr, D., & Chung, G. K. W. K. (2012). *Examining feedback in an instructional video game using process data and error analysis* (CRESST Report 817). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75,* 581–612.

Cai, L., & Houts, C. R. (2021). Longitudinal analysis of patient-reported outcomes in clinical trials: Applications of multilevel and multi-dimensional item response theory. *Psychometrika, 86*, 754–777.

Cai, L., Choi, K., & Kuhfeld, M. (2016). On the role of multilevel item response models in multisite evaluation studies for serious games. In H. F. O'Neil, E. L. Baker, & R. S. Perez (Eds.), *Using games and simulations for teaching and assessment: Key issues* (pp. 280–301). Routledge. https://doi.org/10.4324/9781315817767

Center for Advanced Technology in Schools. (2012). *CATS developed games* (CRESST Resource Report No. 15). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Center for Advanced Technology in Schools. (2013a). *CATS knowledge and item specifications: Functions*. University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Center for Advanced Technology in Schools. (2013b). *CATS knowledge and item specifications: Rational number equivalence*. University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Center for Advanced Technology in Schools. (2013c). *CATS knowledge and item specifications: Solving equations*. University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chi, M. T., Glaser, R., & Farr, M. J. (1988). *The nature of expertise*. Psychology Press.

Choi, K., Parks, C. B., Feng, T., Redman, E. J. K. H., & Chung, G. K. W. K. (2021). *Molly of Denali analytics validation study final report* (Deliverable to PBS KIDS). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Choi, K., Suh, Y. S., Chung, G. K. W. K., Redman, E. J. K. H., Feng, T., & Parks, C. B. (2021). *A secondary analysis of the Molly of Denali RCT data: Examining the relationship among game-based indicators, video usage, and external outcomes using advanced psychometric modeling and population data* (Deliverable to EDC). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K. (2015). Guidelines for the design, implementation, and analysis of game telemetry. In C. S. Loh, Y. Sheng, & D. Ifenthaler (Eds.), *Serious games analytics: Methodologies for performance measurement, assessment, and improvement* (pp. 59–79). Springer.

Chung, G. K. W. K., & Baker, E. L. (2003). An exploratory study to examine the feasibility of measuring problem-solving processes using a click-through interface. *Journal of Technology, Learning, and Assessment, 2*(2). http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1662

Chung, G. K. W. K., Baker, E. L., Vendlinski, T. P., Buschang, R. E., Delacruz, G. C., Michiuye, J. K., Wainess, R., & Bittick, S. J. (2010, April 30–May 4). Testing instructional design variations in a prototype math game. In R. Atkinson (Chair), *Current perspectives from three national R&D centers focused on game-based learning: Issues in learning, instruction, assessment, and game design* [Structured poster session]. Annual meeting of the American Educational Research Association, Denver, CO, United States.

Chung, G. K. W. K., Choi, K., Baker, E., & Cai, L. (2014). *The effects of math video games on learning: A randomized evaluation study with innovative impact estimation techniques* (CRESST Report 841). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., de Vries, L. F., Cheak, A. M., Stevens, R. H., & Bewley, W. L. (2002). Cognitive process validation of an online problem solving assessment. *Computers in Human Behavior*, *18*, 669–684.

Chung, G. K. W. K., Delacruz, G. C., de Vries, L. F., Kim, J.-O., Bewley, W. L., de Souza e Silva, A. A., Sylvester, R. M., & Baker, E. L. (2004). *Determinants of rifle marksmanship performance: Predicting shooting performance with advanced distributed learning assessments* (Deliverable to Office of Naval Research). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., Delacruz, G. C., Dionne, G. B., Baker, E. L., Lee, J. J., & Osmundson, E. (2016). *Towards individualized instruction with technology-enabled tools and methods* (CRESST Report 854). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., Delacruz, G. C., Dionne, G. B., & Bewley, W. L. (2003). Linking assessment and instruction using ontologies. *Proceedings of the I/ITSEC*, *25*, 1811–1822.

Chung, G. K. W. K., & Feng, T. (2024). From clicks to constructs: An examination of validity evidence of game-based indicators derived from theory. In M. Sahin & D. Ifenthaler (Eds.), *Assessment analytics in education. Advances in analytics for learning and teaching* (pp. 327–354). Springer. https://doi.org/10.1007/978-3-031-56365-2_17

Chung, G. K. W. K., Madni, A., & Baker, E. L. (2015). *EAITS test and evaluation final report* (Deliverable to Raytheon BBN). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., Madni, A., Iseli, M., Koenig, A., & Baker, E. L. (2014). *CRESST Assessment report: Validation of knowledge probe methodology* (Deliverable to Raytheon BBN). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., Michiuye, J. K., Brill, D. G., Sinha, R., Saadat, F., de Vries, L. F., Delacruz, G. C., Bewley, W. L., & Baker, E. L. (2002). *CRESST human performance knowledge mapping system* (Final deliverable to the Office of Naval Research). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., Nagashima, S. O., Espinosa, P. D., Berka, C., & Baker, E. L. (2008). *An exploratory investigation of the effect of individualized computer-based instruction on rifle marksmanship performance and skill* (CRESST Tech. Rep. No. 754). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., O'Neil, H. F., Delacruz, G. C., & Bewley, W. L. (2005). The role of affect on novices' rifle marksmanship performance. *Educational Assessment*, *10*, 257–275.

Chung, G. K. W. K., & Parks, C. (2015). *Bundle 1 computational model–v1 (Measurement)* (Deliverable to PBS KIDS). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., Parks, C. B., Redman, E. J. K. H., Choi, K., Kim, J., Madni, A., & Baker, E. L. (2016). *PBS KIDS final report* (Deliverable to PBS KIDS). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chung, G. K. W. K., Redman, E. H., & Madni, A. (2018, January 4–7). *Computer-based assessment of nonroutine problem solving* [Presentation]. Sixteenth Annual Hawaii International Conference on Education, Honolulu, HI, United States.

Chung, G. K. W. K., & Roberts, J. (2018, April 13–17). Common learning analytics for learning games. In E. L. Baker (Chair), *Games and simulations: Learning analytics and metrics* [Symposium]. Annual meeting of the American Educational Research Association, New York, NY, United States.

Chung, G. K. W. K., Ruan, Z., & Redman, E. J. K. H. (2021, April 9–12). *A qualitative comparison of young children's performance on analogous digital and hands-on tasks: Assessment implications* [Paper presentation]. Annual meeting of the American Educational Research Association, Virtual Conference, United States.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* Springer Science & Business Media.

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, *33*(8), 620–639.

Delacruz, G. C. (2011). *Games as formative assessment environments: Examining the impact of explanations of scoring and incentives on math learning, game performance, and help seeking* (CRESST Report 796). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

DiCerbo, K. E., Mislevy, R. J., & Behrens, J. T. (2016). Inference in game-based assessment. In H. F. O'Neil, E. L. Baker, & R. S. Perez (Eds.), *Using games and simulations for teaching and assessment: Key issues* (pp. 253–279). Routledge. https://doi.org/10.4324/9781315817767

Education Development Center, Inc., & SRI International. (2021). *Mahsi'choo for the Info! Molly of Denali teaches children about informational text*. https://www.edc.org/sites/default/files/uploads/EDC-SRI-Mahsi-choo-Info-Summary.pdf

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist*, *49*(8), 725–747.

Espinosa, P. D., Nagashima, S. O., Chung, G. K. W. K., Parks, D., & Baker, E. L. (2009). *Development of sensor-based measures of rifle marksmanship skill and performance* (CRESST Tech. Rep. No. 756). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Feng, T. (2019, April 5–9). *Using game-based measures to assess children's scientific thinking about force* [Poster presentation]. Annual meeting of the American Educational Research Association, Toronto, Canada.

Feng, T., & Cai, L. (2024). Sensemaking of process data from evaluation studies of educational games: An application of cross-classified item response theory modeling. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12396

Feng, T., & Cai, L. (2025). Integrating data from multiple sources in evaluation studies of educational games: An application of cross-classified item response theory modeling. In J. L. Plass & X. Ochoa (Eds.), *Serious games* (Vol. 15259, pp. 70–76). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-74138-8_6

Feng, T., & Chung, G. K. W. K. (2022, April 22–25). Extracting debugging indicators based on distance to solution in a block-based programming game. In G. K. W. K. Chung (Chair), *Game-based indicators of learning processes: Extraction methods, validity evidence, and applications* [Symposium]. Annual meeting of the American Educational Research Association, San Diego, CA, United States.

Feng, T., Chung, G. K. W. K., Choi, K., & Redman, E. J. K. H. (2025). *EDC SRI Wombats secondary analysis–Final report* (Deliverable to PBS KIDS). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Fitts, P. M., & Posner, M. I. (1967). *Human performance*. Brooks/Cole.

Gates, A. I. (1918). The abilities of an expert marksman tested in the psychological laboratory. *Journal of Applied Psychology, 2,* 1–14.

Gentner, D., & Gentner, D. R. (1983). Flowing waters or teeming crowds: Mental models of electricity. In D. Gentner, & A. L. Stevens (Eds.), *Mental models* (pp. 99–129). Psychology Press.

Gómez, M. J., Ruipérez-Valiente, J. A., & Clemente, F. J. G. (2022). A systematic literature review of game-based assessment studies: Trends and challenges. *IEEE Transactions on Learning Technologies,* 1–16. https://doi.org/10.1109/TLT.2022.3226661

Gorman, T. P., Purves, A. C., & Degenhart, R. E. (1988). *The IEA study of written composition I: The international writing tasks and scoring scales.* Pergamon.

Graves, D. (1978). *Balance the basics: Let them write.* Ford Foundation.

Gris, G., & Bengtson, C. (2021). Assessment measures in game-based learning research: A systematic review. *International Journal of Serious Games*, *8*(1), Article 1. https://doi.org/10.17083/ijsg.v8i1.383

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies, 43*(5), 907–928. https://doi.org/10.1006/ijhc.1995.1081

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items* (1st ed.). Routledge. https://doi.org/10.4324/9780203850381

Herl, H. E., Baker, E. L., & Niemi, D. (1996). Construct validation of an approach to modeling cognitive structure of US history knowledge. *Journal of Educational Research*, *89*(4), 206–218.

Herl, H. E., O'Neil Jr, H. F., Chung, G. K., & Schacter, J. (1999). Reliability and validity of a computer-based knowledge mapping system to measure content understanding. *Computers in Human Behavior*, *15*(3–4), 315–333.

Ihlenfeldt, S., Chung, G. K. W., K., Lyons, S., Lawson, J., & Redman, E. J. K. H. (2025). *Modeling the relationships among online Solitaire gameplay and measures of cognition* (CRESST Report 877). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Iseli, M. R., & Jha, R. (2016). Computational issues in modeling user behavior in serious games. In H. F. O'Neil, E. L. Baker, & R. S. Perez (Eds.), *Using games and simulations for teaching and assessment: Key issues* (pp. 21–40). Routledge.

Iseli, M. R., Koenig, A. D., Lee, J. J., & Wainess, R. (2010). *Automated assessment of complex task performance in games and simulations* (CRESST Report 775). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Iseli, M. R., Lee, J. J., Schenke, K., Leon, S., Lim, D., Jones, B., & Cai, L. (2019). *Simulation-based assessment of ultrasound proficiency* (CRESST Report 865). UCLA/CRESST

Jiao, H., He, Q., & Veldkamp, B. P. (2021). Editorial: Process data in educational and psychological measurement. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.793399

Kerr, D. (2014). *Identifying common mathematical misconceptions from actions in educational video games* (CRESST Report 838). UCLA/CRESST

Kerr, D., & Chung, G. K. W. K. (2012a). Identifying key features of student performance in educational video games and simulations through cluster analysis. *Journal of Educational Data Mining, 4*, 144–182.

Kerr, D., & Chung, G. K. W. K. (2012b). *Using cluster analysis to extend usability testing to instructional content* (CRESST Report 816). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kerr, D., & Chung, G. K. W. K. (2013a). Identifying learning trajectories in an educational video game. In R. Almond & O. Mengshoel (Eds.), *Proceedings of the 2013 UAI Application Workshops: Big Data Meet Complex Models and Models for Spatial, Temporal and Network Data* (pp. 20–28). http://ceur-ws.org/Vol-1024/

Kerr, D., & Chung, G. K. W. K. (2013b). *The effect of in-game errors on learning outcomes* (CRESST Report 835). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kim, Y. J., & Ifenthaler, D. (2019). Game-based assessment: The past ten years and moving forward. In D. Ifenthaler & Y. J. Kim (Eds.), *Game-based assessment revisited: Advances in game-based learning* (pp. 3–11). Springer. https://doi.org/10.1007/978-3-030-15569-8_1

Koenig, A. D., Lee, J. J., Iseli, M., & Wainess, R. (2010). *A conceptual framework for assessing performance in games and simulation* (CRESST Report 771). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kumar, R., Chung, G. K. W. K., Madni, A., & Roberts, B. (2015). First evaluation of the physics instantiation of a problem-solving based online learning platform. In C. Conati, N. Hefferman, A. Mitrovic, & M. F. Verdejo (Eds.), *Lecture Notes in Computer Science: Vol. 9112. Artificial Intelligence in Education* (pp. 686–689). Springer.

Landers, R. (2015). Special issue on assessing human capabilities in video games and simulations. *International Journal of Gaming and Computer-Mediated Simulations, 7*(4), iv–viii.

Levy, R. (2019). Dynamic Bayesian network modeling of game-based diagnostic assessments. *Multivariate Behavioral Research*, *54*(6), 771–794. https://doi.org/10.1080/00273171.2019.1590794

Lindner, M. A., & Greiff, S. (2023). Process data in computer-based assessment. *European Journal of Psychological Assessment*, *39*(4), 241–251. https://doi.org/10.1027/1015-5759/a000790

Madni, A., Griffin, N., & Yang, J. S. (2013, April 27–May 1). *Integrating assessment of SEL into an early childhood science learning context* [Paper presentation]. Annual meeting of the American Educational Research Association, San Francisco, CA, United States.

Mislevy, R. J., Corrigan, S., Oranje, A., DiCerbo, K., Bauer, M. I., von Davier, A., & John, M. (2015). Psychometrics and game-based assessment. In F. Drasgow (Ed.), *Technology and testing* (pp. 23–48). Routledge. https://doi.org/10.4324/9781315871493

Nagashima, S. O., Chung, G. K. W. K., Espinosa, P. D., & Berka, C. (2009). Sensor-based assessment of basic rifle marksmanship. *Proceedings of the I/ITSEC*, Orlando, FL.

Niemi, D., & Baker, E. L. (1998, January). *Design and development of a comprehensive assessment system: Pilot testing, scoring, and refinement of mathematics and language arts performance assessments* (Final Deliverable). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

O'Neil, H. F. Jr, & Baker, E. L. (1987). *Issues in intelligent computer-assisted instruction, evaluation and measurement*. In T. B. Gutkin & S. L. Wise (Eds.), *The computer and decision-making process* (pp. 199–224). Erlbaum.

O'Neil, H. F., Baker, E. L., & Linn, R. L. (1990, April 16–20). *Performance assessment framework* [Paper presentation]. Annual meeting of the American Educational Research Association, Boston, MA, United States.

O'Neil, H. F., Baker, E. L., Ni, Y., Jacoby, A., & Swigger, K. M. (1994). *Human benchmarking for the evaluation of expert systems.* In H. F. O'Neil, Jr., & E. L. Baker (Eds.), Technology assessment in software applications (pp. 13–45). Lawrence Erlbaum Associates.

O'Neil, H. F., Mayer, R. E., Rueda, R., & Baker, E. L. (2021). Measuring and increasing self-efficacy in a game. In H. F. O'Neil, E. L. Baker, R. S. Perez, & S. E. Watson (Eds.), *Using cognitive and affective metrics in educational simulations and games: Applications in school and workplace contexts* (pp. 131–158). Routledge/ Taylor & Francis.

Oranje, A., Mislevy, B., Bauer, M. I., & Jackson, G. T. (2019). Summative game-based assessment. In D. Ifenthaler & Y. J. Kim (Eds.), *Game-based assessment revisited* (pp. 37–65). Springer. https://doi.org/10.1007/978-3-030-15569-8_3

Organisation for Economic Co-operation and Development (OECD). (2014). *PISA 2012 Results: Creative problem solving: Students' skills in tackling real-life problems* (Volume V). OECD Publishing. http://dx.doi.org/10.1787/9789264208070-en

Organisation for Economic Co-operation and Development (OECD). (2021). *OECD digital education outlook 2021: Pushing the frontiers with artificial intelligence, blockchain and robots*. https://doi.org/10.1787/589b283f-en

Quellmalz, E. S. (1982). *Designing writing assessments: Balancing fairness, utility, and cost* (CSE Report 188). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). University of Chicago Press.

Redman, E. J. K. H., Chung, G. K. W. K., Feng, T., Parks, C. B., Schenke, K., Michiuye, J. K., Choi, K., Ziyue, R., & Wu, Z. (2020). *Cat in the Hat Builds That analytics validation study—final deliverable* (Deliverable to PBS KIDS). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Redman, E. J. K. H., Chung, G. K. W. K., Schenke, K., Maierhofer, T., Parks, C. B., Chang, S. M., Feng, T., Riveroll, C. S., & Michiuye, J. K. (2018). *Connected learning final report* (Deliverable to PBS KIDS). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Redman, E. J. K. H., Feng, T., Parks, C. B., Choi, K., & Chung, G. K. W. K. (2023). *Learning-related analytics KPI—KPI final report* (Deliverable to PBS KIDS). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Redman, E. J. K. H., Feng, T., Parks, C. B., Chung, G. K. W. K., Choi, K., & Cai, L. (2025). *Wombats analytics evaluation –Final report* (Deliverable to PBS KIDS). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Redman, E. J. K. H., Parks, C. B., Michiuye, J. K., Suh, Y. S., Chung, G. K. W. K., Kim, J., & Griffin, N. (2021). *Social-emotional learning games validity study (exploratory study): Final study report*. University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Richardson, M. W., Russell, J. T., Stalnaker, J. M., & Thurstone, L. L. (1933). *Manual of examination methods*. The University of Chicago. http://hdl.handle.net/2027/uiug.30112066775344

Roberts, J. D., Chung, G. K. W. K., & Parks, C. B. (2016). Supporting children's progress through the PBS KIDS learning analytics platform. *Journal of Children and Media, 10*, 257–266.

Roberts, J. D., Younger, J. W., Corrado, K., Felline, C., & Lovato, S. (in press). Practical examples of assessment in the service of learning at PBS KIDS. In Tucker, E., Everson, E., Baker, E. L., & E. W. Gordon (Eds.), *Handbook for assessment in the service of learning, Volume III, Examples of assessment in the service of learning*. University of Massachussets Amherst.

Ruiz-Primo, M. A., Shavelson, R. J., Li, M., & Schultz, S. E. (2001). On the validity of cognitive interpretations of scores from alternative concept-mapping techniques. *Educational Assessment*, *7*(2), 99–141.

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

Savitsky, E. (2013). *U.S. Patent No. 8,480,404*. U.S. Patent and Trademark Office.

Scardamalia, M., Bereiter, C., & Steinbach, R. (1984). Teachability of reflective processes in written composition. *Cognitive Science, 8*(2), 173–190.

Schacter, J., Herl, H. E., Chung, G. K. W. K., Dennis, R. A., & O'Neil, H. F. (1999). Computer-based performance assessments: A solution to the narrow measurement and reporting of problem-solving. *Computers in Human Behavior*, *15*, 403–418.

Shute, V., & Wang, L. (2016). Assessing and supporting hard-to-measure constructs in video games. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment* (pp. 535–562). John Wiley & Sons. https://doi.org/10.1002/9781118956588.ch22

Sireci, S. G. (2016). Commentary on chapters 1–4: Using technology to enhance assessments. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 104–108). Routledge. https://doi.org/10.4324/9781315871493

Stevens, R., Ikeda, J., Casillas, A., Palacio-Cayetano, J., & Clyman, S. (1999). Artificial neural network-based performance assessments. *Computers in Human Behavior*, *15*(3), 295–313. https://doi.org/10.1016/S0747-5632(99)00025-4

Tadayon, M., & Pottie, G. J. (2020). Predicting student performance in an educational game using a hidden Markov model. *IEEE Transactions on Education, 63*(4), 299–304.

Tlili, A., Chang, M., Moon, J., Liu, Z., Burgos, D., Chen, N.-S., & Kinshuk. (2021). A systematic literature review of empirical studies on learning analytics in educational games. *International Journal of Interactive Multimedia and Artificial Intelligence, 7*(2), 250–261. http://doi.org/10.9781/ijimai.2021.03.003

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2015, November). *WWC review of the report: The effects of math video games on learning: A randomized evaluation study with innovative impact estimation techniques.* http://whatworks.ed.gov

U.S. Marine Corps. (2001). *Rifle marksmanship* (PCN 144 000091 00, MCRP 3–01A).

VanLehn, K., Chung, G., Grover, S., Madni, A., & Wetzel, J. (2016). Learning science by constructing models: Can Dragoon increase learning without increasing the time required? *International Journal of Artificial Intelligence in Education, 26*, 1033-1068. https://doi.org/10.1007/s40593-015-0093-5

Vendlinski, T. P., Chung, G. K. W. K., Binning, K. R., & Buschang, R. E. (2011). *Teaching rational number addition using video games: The effects of instructional variation.* (CRESST Report 808). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Wigger, K. M., O'Neil, H. F., Jr., & Ni, Y. (1990). *Evaluation of expert systems: A review of the literature.* Los Angeles: University of California, Center for the Study of Evaluation/Center for Technology Assessment. https://apps.dtic.mil/sti/tr/pdf/ADA233601.pdf

Wiley, K., Robinson, R., & Mandryk, R. L. (2021). The making and evaluation of digital games used for the assessment of attention: Systematic review. *JMIR Serious Games, 9*(3), e26449. https://doi.org/10.2196/26449

Zumbo, B. D., Maddox, B., & Care, N. M. (2023). Process and product in computer-based assessments: Clearing the ground for a holistic validity framework. *European Journal of Psychological Assessment*, *39*(4), 252–262. https://doi.org/10.1027/1015–5759/a00074.