FROM CLICKS TO CONSTRUCTS: AN EXAMINATION OF VALIDITY EVIDENCE OF GAME-BASED INDICATORS DERIVED FROM THEORY

Tianying (Teanna) Feng
CRESST / University of California, Los Angeles

Gregory K.W.K. Chung
CRESST / University of California, Los Angeles

Paper presented at the annual meeting of the
American Educational Research Association
San Diego, CA
April 22, 2022

University of California, Los Angeles (UCLA)
School of Education & Information Studies (SEIS)
National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
300 Charles E. Young Drive, North
GSE&IS Building, Room 101D
Los Angeles, CA 90095-1522
cresst.org

From Clicks to Constructs: An Examination of Validity Evidence of Game-Based Indicators Derived From Theory[1]

ABSTRACT

A critical issue in using fine-grained gameplay data to measure learning processes is the development of indicators and the algorithms used to derive such indicators. Successful development—that is, developing traceable, interpretable, and sensitive-to-learning indicators—requires understanding the underlying theory, how the theory is instantiated under different game conditions, and programming that considers these exigencies using fine-grained gameplay data. This study examines preliminary validity evidence on various game-based indicators (GBIs) developed from theoretical frameworks and not from unsupervised feature extraction methods. Examples of indicators are presented that represent three types of indicators: Common indicators, distance-based indicators, and games-specific indicators. For each example, the theoretical background is presented briefly, followed by a description of how the indicator design flows from the theory. Correlational analyses show how the GBIs relate to external criterion measures. Limitations and next steps are discussed.

INTRODUCTION

The interest in using games **for measurement purposes** lies in the promise that games, compared to traditional testing formats, offer engaging ways to more comprehensively measure learners' knowledge, skills, and abilities. Games may heighten engagement and sustain learners' interest and attention, thereby allowing observation of effortful performance over a prolonged time and under various conditions, allowing for a richer sampling of behavior. Finally, games can be instrumented to collect fine-grained, moment-to-moment observations of learners' in-game choices and actions yoked to changes in the state of the game (for a discussion of these issues, see Baker et al., 2011; Baker & Delacruz, 2016; DiCerbo & Behrens, 2012; Landers, 2015; Mislevy et al., 2016; Oranje et al., 2019; Sireci, 2016; Shute & Wang, 2016). Indeed, a common aspirational goal is to "replace the dull, time-consuming, and anxiety-producing traditional approaches commonly used today" (Landers, 2015, p. vii). While Landers was referring to traditional standardized tests, his sentiment reflects the general desire to develop other means of measuring what learners know and can do under more engaging and complex situations. This idea is so compelling, that McKinsey & Company, the premier consulting

[1] The research reported here was supported by awards to UCLA by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A190433 to the University of California, Los Angeles, and other grants from the U.S. Department of Education (PR/Award Nos. U295A 150003 and S295A200004, CFDA No. 84.295A]). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The contents of this report do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government. Additional support for this work was provided by Silicon Valley Community Foundation. However, the content of this report do not necessarily represent the policy of the Silicon Valley Community Foundation, and you should not assume endorsement by the Silicon Valley Community Foundation.

company in North America, has included a game-based problem-solving assessment as part of their recruitment process (2021a, 2021b).

In comparison to traditional test formats (e.g., multiple-choice tests) or even modern technology-based assessment tasks (e.g., NAEP [Bennett et al., 2007; NCES, 2012], PISA [OECD, 2014, 2017]), games can elicit much more varied responses from players: Player actions can be repetitive, erratic, or systemic; players respond differently to feedback and use game resources in different ways; and players can directly affect the state of the game as well as be affected by the game. Reasoning and learning unfold over time, and gameplay behavior is highly variable both between individuals and within individuals. Furthermore, the design of games themselves—what makes a game a game—often encourages variation in gameplay, including branching where not all players encounter the same levels; player choice of what levels to play; no limit on the time spent on a level; multiple attempts at beating a level with explicit and implicit feedback given; levels that can be replayed; and gameplay where failure is expected and makes the game engaging.

To be taken seriously as assessments, however, games need to be subjected to the same rigor as traditional test items. Progress during the last decade has focused on integrating game design and assessment design (i.e., aligning assessment goals, game design and game mechanics, and indicators of competency) and far less on gathering validity evidence on how the indicators generated from gameplay related to learning outcomes, or on describing the algorithm or method underlying the computation of GBIs (Gris & Bengtson, 2021; Kim & Ifenthaler, 2019). This paper focuses on the latter two aspects.

Definitions

By ***telemetry*** (or log data or clickstream), we mean time-stamped packets of data that encapsulate a specific learner action (e.g., adding a command) and related parameters. Well-designed telemetry specifies what learner- and system-initiated events and system states to record and has a well-defined structure (Chung, 2015). By game-based ***indicator,*** we mean an observed value of some variable used directly, as an input to auxiliary transformations to derive secondary indicators, or as an input to a statistical model or analysis procedure. GBIs are extracted from telemetry. By ***traceable,*** we mean one can inspect the algorithm (or code) and trace how the raw telemetry is processed and transformed (e.g., filtered, aggregated, recombined) into a value. By ***interpretable,*** we mean given an algorithm, one can agree or disagree with the meaning ascribed to the variable in light of the assumptions, constraints, and transformations encoded in the algorithm. By ***generalizable,*** we mean an algorithm based on a theory that presumably encodes the rules and conditions described or predicted by the theory and (with modification) can apply to games that may differ in format, mechanics, content, or even learning goals.

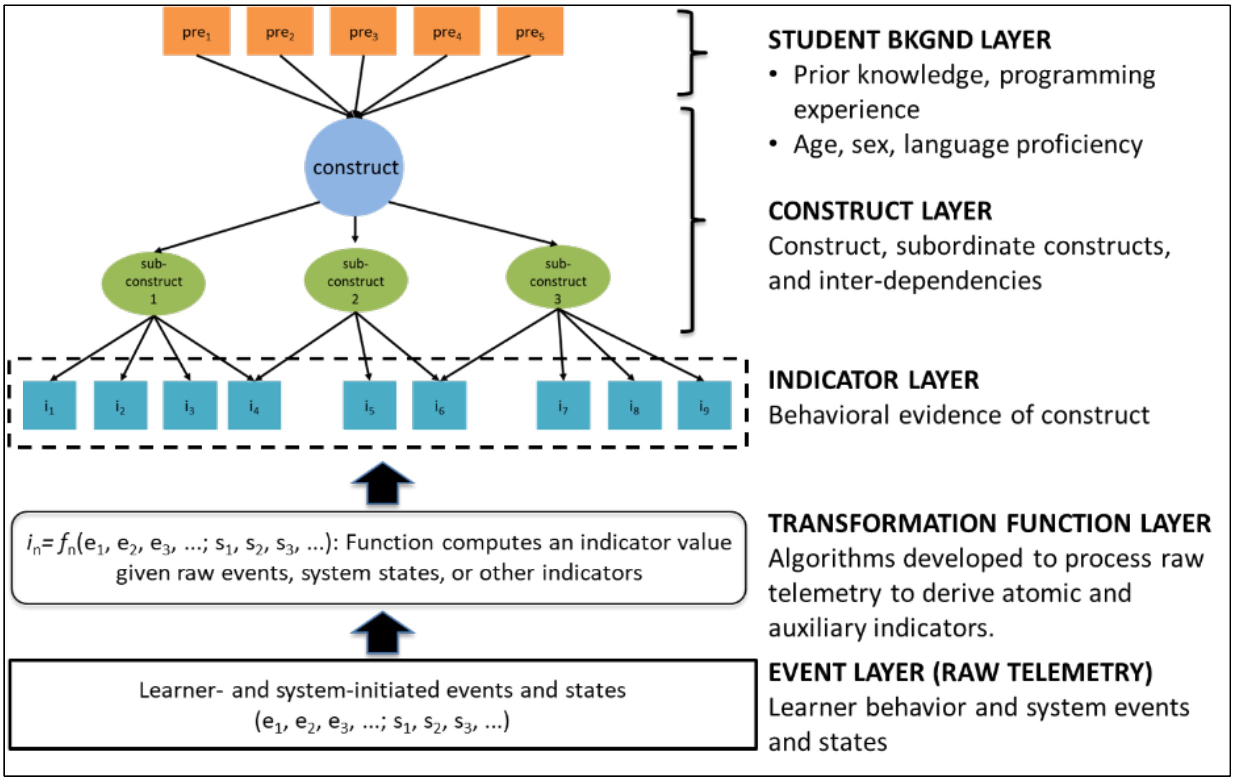
Game-based Indicators (GBI)

A critical issue in measuring learning processes is the indicators: What is their definition, what constructs do they purport to indicate, and what is the empirical evidence supporting its interpretation? Often unreported in literature is the algorithm used to transform raw gameplay telemetry (i.e., log data) into indicators. Indicator identification is often relegated to machine learning procedures. Machine learning procedures may identify variables highly predictive of some criteria, but when the actual variables used in the prediction equation are inspected, the explanatory power of that variable in relation to the claims about what the equation means may suffer. For example, the machine learning algorithm used in a very popular science simulator to detect "designing controlled experiments" uses time a student spent performing various actions related to variable manipulation in the simulation on a science experiment task. Time was used simply because it yielded the best classification performance. It is difficult to explain how differences in minutes determine whether a student was engaged in "designing controlled experiments." (Sao Pedro et al., 2013).

An alternative approach is to develop algorithms guided by theory. Some advantages of starting from theory include algorithms and indicators that are more likely to be traceable, interpretable, fit into a theoretical framework from which it is derived, more generalizable across games, and be sensitive to learning. Our general approach is illustrated in the conceptual model shown in Figure 1. Figure 1 shows how raw telemetry is successively transformed into indicators that can serve as inputs to psychometric and statistical models or other analysis procedures—that is, going from clicks to constructs.

Figure 1

Conceptual Model of the Relations Between Telemetry, Algorithms, and Indicators



pre_1
pre_2
pre_3
pre_4
pre_5
construct
sub-construct 1
sub-construct 2
sub-construct 3
i_1
i_2
i_3
i_4
i_5
i_6
i_7
i_8
i_9
$i_n = f_n(e_1, e_2, e_3, \ldots; s_1, s_2, s_3, \ldots)$: Function computes an indicator value given raw events, system states, or other indicators
Learner- and system-initiated events and states
$(e_1, e_2, e_3, \ldots; s_1, s_2, s_3, \ldots)$
STUDENT BKGND LAYER
• Prior knowledge, programming experience
• Age, sex, language proficiency
CONSTRUCT LAYER
Construct, subordinate constructs, and inter-dependencies
INDICATOR LAYER
Behavioral evidence of construct
TRANSFORMATION FUNCTION LAYER
Algorithms developed to process raw telemetry to derive atomic and auxiliary indicators.
EVENT LAYER (RAW TELEMETRY)
Learner behavior and system events and states

Research Question

Our research question **is, to what extent do game-based indicators relate to criterion measures of learning?** To address this question, we summarize preliminary validity evidence on various game-based indicators (GBI) developed for a variety of games across numerous studies, where the GBIs represent learning outcomes, processes, or strategies. An essential feature of this work is that all GBIs were developed from a theoretical framework and not from unsupervised feature extraction methods.

In the remainder of this paper, we first present the general methods we have used to derive GBIs. We then present five examples drawn from prior work to illustrate how we used theory to guide the development of GBIs. For each example, we present results from correlational analyses of how the GBIs relate to an external criterion measure.

METHOD

General Methods Used to Derive GBIs

In this section, we describe the general methods used to compute three general types of GBIs: (a) “common” indicators, (b) dissimilarity-based indicators, and (c) game-specific strategy indicators.

One type of GBI is summary indicators from gameplay data that presumably reflect the use of the target construct or skill. For example, indicators that decompose gameplay processes into progress- and performance-related outcomes, such as accuracy (performance) and time (progress). A second type of indicator can be used when a product or structure is an outcome of gameplay (e.g., programming code or a virtual structure). The player’s solution structure can be compared to the referent structure to compute a dissimilarity index. A third method is specific to a particular game and requires translating gameplay behavior that maps to qualitative observations or other findings from prior research.

Common Indicators

Common indicators are outcomes that can be easily computed from ‘typical’ gameplay data to provide a general summary of gameplay. Table 1 shows a list of common indicators from prior studies (Chung & Parks, 2015a, 2015b; Chung & Roberts, 2018).

The rationale for developing common indicators is twofold. First, if common indicators can be defined and operationalized, then those indicators may be used to compare different games. One such comparison is for measurement purposes—to identify games that can serve as measures of a child’s proficiency on the particular topic targeted by the game.

The second reason is for practical purposes. Developing game-specific indicators, as well as indicators describing gameplay processes and patterns, can be challenging and require complex data wrangling and coding to transform fine-grained telemetry into more abstract indicators that reflect learning. The complexity of the programming is governed by the telemetry structure, the availability of game state information, and the amount of transformation needed. In general, nontrivial amounts of programming effort are required with extensive exploratory analysis to refine the measure.

Table 1

List of Common Indicators Based on Prior Work (Chung & Parks, 2015a, 2015b; Chung & Roberts, 2018)

Type	Indicator	Descriptions
Performance	Number of correct first attempts	Across gameplay sessions, the number of times players submit or explicitly indicate their answer (e.g., by pressing the submit button).
	Number of total attempts	Across gameplay sessions, the number of times players submit or explicitly indicate their answer (e.g., by pressing the submit button).
	Number of correct attempts	Across gameplay sessions, the number of times players submit or explicitly indicate their answer and reach the goal.
	Number of incorrect attempts	Across gameplay sessions, the number of times players submit or explicitly indicate their answer and fail to reach the goal.
Progress	Mean time spent per level	Mean time spent playing each level across gameplay sessions.
	Total time spent in the game	Cumulative summation of the amount of time spent in the game across gameplay sessions.
	Number of levels completed	Cumulative summation of the number of levels completed across gameplay sessions. Repeated levels are included.
	Number of distinct levels played	Cumulative summation of the number of non-repeated levels played across gameplay sessions.

Note. We use the term "game session" to refer to a pre-defined time point or period of gameplay.

Dissimilarity-Based Indicators

Dissimilarity-based indicators consider the dissimilarity (or distance) between players' responses and a referent. The dissimilarity indicator reflects how far a player's responses are from an optimal state conceptually (e.g., distance to the solution) or graphically (e.g., Euclidean distance computed using x-y coordinates). The optimal state can be a solution, a goal state, or an in-game target object.

If a solution is used as the referent, the referent can be used in at least three ways: (a) as an absolute reference—a gold standard solution(s); (b) as a within-player reference—the same player's last successful attempt becomes the referent, or (c) between-player (and sample-dependent) reference—based on all players' solution attempts, use the attempt that is closest to the actual solution as the reference.

We can also compute dissimilarities with different levels of granularity, with the most fine-grained level being players' moment-to-moment moves. For example, if we compute a dissimilarity value whenever players make a move (e.g., a click), and multiple moves can be nested within an attempt or submission, then the dissimilarity indicator itself is a time series. An alternative is to derive the indicator by considering only the final submission for each game level.

Secondary indicators can be derived using the dissimilarity information, such as whether there are diminishing differences between the player's submissions and the solution (i.e., convergence over time) or how fast the convergence is. In this case, we have the option to model the process as measured by such differences directly or use secondary indicators as scored versions of players' performance and process.

Game-specific Indicators

Game-specific indicators are dependent on the particular game. This type of indicator may be a "one-off."

RESULTS

Example GBIs and Results

We next present several examples that illustrate using one of the methods to derive GBIs. For each example, we note which general method was used, the theoretical rationale behind the operationalization of the indicator, and preliminary validity evidence of how well the GBI related to independent measures of the target construct or skill.

Example 1: Speed and Accuracy

Target knowledge or skill:	Numeracy
Method used to derive the GBI:	Common indicators
Theoretical justification:	Skill mastery can be described by performance accuracy and speed of task completion, a robust finding across motor and verbal domains (e.g., Ackerman, 1990; Anderson, 1982; Fitts & Posner, 1967).
Validation approach:	Correlate GBI to external criterion measure (TEMA3)
Summary of results:	• Performance indicators Knowledge of numeracy: (a) positively related to performance in the game (ρ ranges from .4 to .6, $p < .01$ for correct first attempt at a solution, and (b) negatively related to the number of failed solution attempts (ρ ranges from -.4 to -.6, $p < .01$ for number of incorrect solution attempts). • Progress indicators (*MeatBall Launcher*) Knowledge of numeracy: (a) positively related to progress in the game ($\rho = .67$, $p < .001$), (b) negatively related to the amount of time taken to beat a level ($\rho = -.57$, $p < .001$).
Original study or source of data:	Chung, G. K. W. K., & Parks, C. (2015). *Bundle 1 computational model analysis report* (Deliverable to PBS KIDS). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Theoretical background. This example of GBIs is based on the pervasive use of speed and accuracy to describe human performance. That is, skill mastery can be characterized by performance accuracy and speed of task completion, a robust finding across motor and verbal domains (e.g., Ackerman, 1990; Anderson, 1982; Fitts & Posner, 1967). We expected these two variables to be important in describing skill attainment but had no basis for asserting which component was more important.

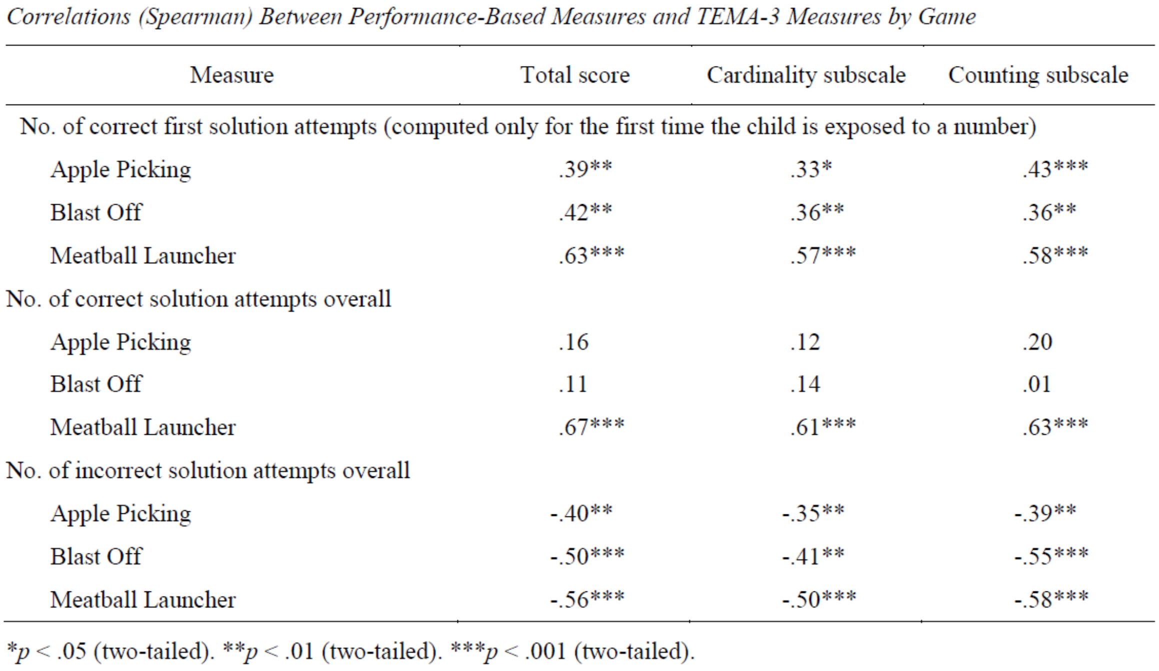
GBI algorithm. We examined three games designed to promote numeracy for preK children. The game mechanics differed by game, with *MeatBall Launcher* being the most "test like" where children were asked to collect the target number of meatballs on a plate and then launch the meatballs onto a plate of spaghetti. Each level had a new target number. No feedback was given before launching the meatballs. Players were told if they had collected an incorrect number of meatballs. The two other games were not as direct in the relation between the game mechanic and the application of the target knowledge.

We created GBIs that represented various aspects of the time component (i.e., game progress) and performance component (e.g., correct solution attempt). Table 1 contains the set of indicators computed.

External criterion measure. Twenty-two items of the Test of Early Mathematics Ability, 3rd Edition (TEMA-3) (Ginsburg & Baroody, 2003) were administered after students completed all games. The items measured numbering, counting and cardinality, and reading numerals, and the KR20 reliability was .91.

Preliminary validity evidence. Table 2 shows the correlations between the GBIs and the external criterion measure (Test of Early Mathematics Ability, 3rd Edition [TEMA-3], Ginsburg & Baroody, 2003). The pattern of (Spearman) correlations suggests that more knowledge of numeracy was related positively to higher performance in the game (ρ ranges from .4 to .6, $p < .01$ for correct first attempt at a solution) and negatively to errors in the game (ρ ranges from -.4 to -.6, $p < .01$ for number of incorrect solution attempts). In terms of progress indicators, only one game (*MeatBall Launcher*) showed significant relationships. Knowledge of numeracy was related to how far in the game players advanced ($\rho = .67$, $p < .001$) and negatively related to how long a player took on average for their first correct attempt ($\rho = -.57$, $p < .001$).

Table 2

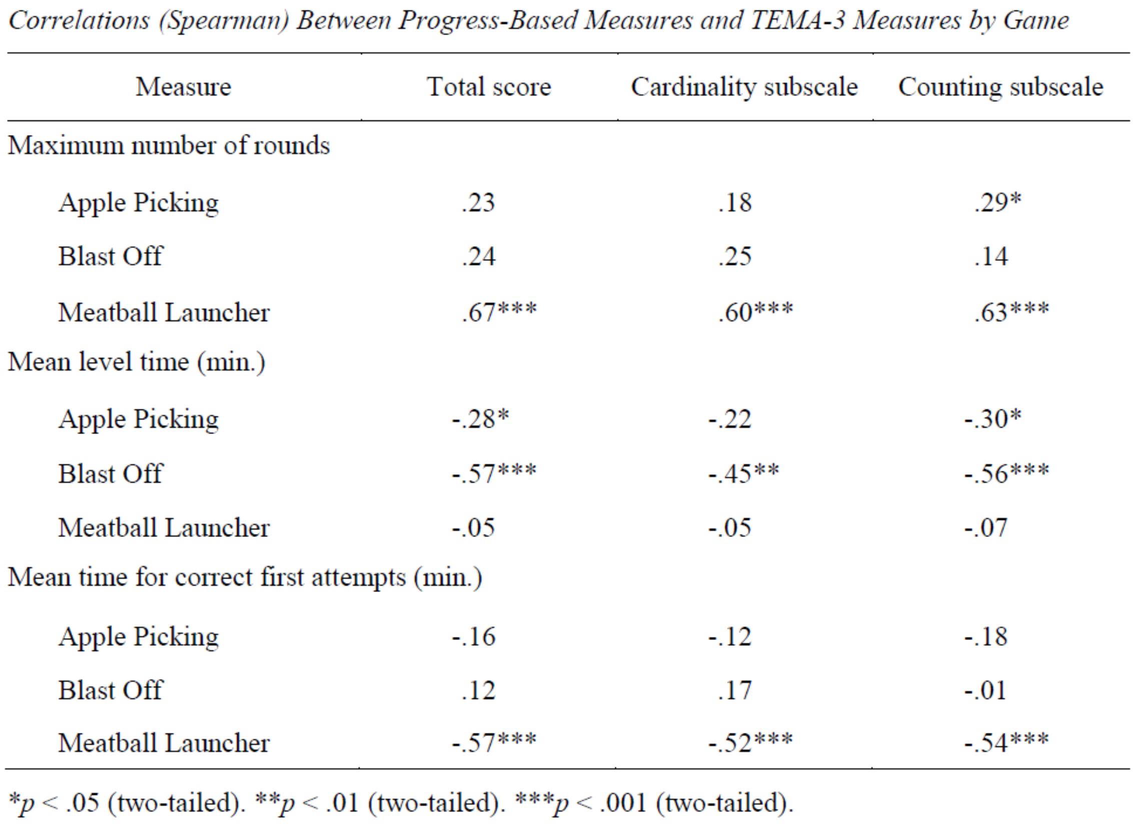


Correlations (Spearman) Between Performance-Based Measures and TEMA-3 Measures by Game

Measure	Total score	Cardinality subscale	Counting subscale
No. of correct first solution attempts (computed only for the first time the child is exposed to a number)			
Apple Picking	.39**	.33*	.43***
Blast Off	.42**	.36**	.36**
Meatball Launcher	.63***	.57***	.58***
No. of correct solution attempts overall			
Apple Picking	.16	.12	.20
Blast Off	.11	.14	.01
Meatball Launcher	.67***	.61***	.63***
No. of incorrect solution attempts overall			
Apple Picking	-.40**	-.35**	-.39**
Blast Off	-.50***	-.41**	-.55***
Meatball Launcher	-.56***	-.50***	-.58***

*p < .05 (two-tailed). **p < .01 (two-tailed). ***p < .001 (two-tailed).

Table 3



Correlations (Spearman) Between Progress-Based Measures and TEMA-3 Measures by Game

Measure	Total score	Cardinality subscale	Counting subscale
Maximum number of rounds			
Apple Picking	.23	.18	.29*
Blast Off	.24	.25	.14
Meatball Launcher	.67***	.60***	.63***
Mean level time (min.)			
Apple Picking	-.28*	-.22	-.30*
Blast Off	-.57***	-.45**	-.56***
Meatball Launcher	-.05	-.05	-.07
Mean time for correct first attempts (min.)			
Apple Picking	-.16	-.12	-.18
Blast Off	.12	.17	-.01
Meatball Launcher	-.57***	-.52***	-.54***

*p < .05 (two-tailed). **p < .01 (two-tailed). ***p < .001 (two-tailed).

Table 11 in the Appendix shows a compilation of results for various games showing correlations between the external outcome used for the game (typically a test) and the time component and performance accuracy indicators.

Example 2: Scientific Thinking

Target knowledge or skill:	Scientific thinking
Method used to derive the GBI:	Dissimilarity-based indicators
Theoretical justification:	Scientific thinking is the ability to evaluate evidence and revise prior beliefs. Hypothesis generation, experimentation, evidence evaluation, and theory revision are all important for successful problem solving, scientific understanding, and inquiry (Kuhn, 2002; Simon, 1977; Zimmerman, 2007).
Validation approach:	Correlate GBIs to an external criterion measure
Summary of results:	• Knowledge of force and motion was related to effective force adjustments (ρ = .23, p < .05), suggesting that players with higher knowledge of force and motion were better able to respond to in-game feedback and adjust the force in the desired direction. • Knowledge of force and motion was positively related to productive game actions (ρ = .23, p < .05) and negatively related to unproductive game actions (ρ = -.21, p < .05).
Original study or source of data:	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). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

Theoretical background. Kuhn (2002) argues that the core of scientific thinking (ST) is the ability to evaluate evidence and revise prior beliefs. Hypothesis generation, experimentation, evidence evaluation, and theory revision are important for successful problem solving, scientific understanding, and inquiry (Kuhn, 2002; Simon, 1977; Zimmerman, 2007). Children as young as kindergarten have been found to develop ST (Gopnik, 2012) and ST processes, such as experimentation and evidence evaluation, predicted young children's domain-specific knowledge (Edelsbrunner, Schalk, Schumacher, & Stern, 2015; van der Graaf, Segers, & Verhoeven, 2018).

GBI algorithm. The game *Fish Force* targeted the concepts of force and motion. In a typical game level, players are presented with several core components: a rink, a launcher that moves along four sides of the rink, a toy, and a target location. The level objective is to move the toy to the target by hitting the toy with a projectile. The projectile's motion is governed by players adjusting the launcher's force, the launcher's position, or both.

When players experimented with how changing the launcher's force may affect the projectile's movement and the outcome, they also used ST processes (e.g., hypothesis generation; Kuhn, 2002; Simon, 1977; Zimmerman, 2007). The indicators derived were designed to identify sequences of ***behaviors that suggest children revising their strategies in response to visual or auditory feedback***. Levels where players succeeded in the first attempt and levels that did not provide specific force feedback were excluded from the analysis.

We focused on the force adjustment and derived GBIs corresponding to two important components of ST—evaluation of evidence (as provided by in-game feedback) and strategy revision (as reflected by subsequent adjustments).

We first developed auxiliary indicators that represented how much players' adjustments of force deviated from a target value. The indicators were: (a) *Force deviation*—the difference (deviation) between the amount of force set by the player and the target amount of force; (b) *Force convergence*—whether each adjustment was converging to the optimal value, reflected by diminishing deviations. The mean values for both, averaged across all game levels played, were used in the final analysis.

Then a secondary indicator was derived to represent the quality of an adjustment, termed *Effective force adjustment,* and intended to meet the core definition of ST—the ability to evaluate evidence and revise prior beliefs. The indicator also was conditioned by children's responses to specific feedback events. The algorithm assumed that children's adjustments were informed by context and were not a random process. Instead, children's responses were responses to feedback events in the game.

Effective force adjustment was computed as the average number of times a child used information from a previous feedback event (evidence) and adjusted the force to a value closer to the solution (theory revision). We expected players who exhibited more effective force adjustments would have better ST and thus better performance on the external assessment.

External criterion measure. The external knowledge assessment consisted of five items developed and pilot tested by Redman et al. (2019). Redman et al.'s (2019) study used a between-subjects pretest-posttest randomized design. The reliability of the assessment as measured by Cronbach's alphas was .64 for the pretest and .57 for the posttest. Because the game was used as a potential intervention in Redman et al. (2019), the posttest sum score should reflect players' most recent standing after gameplay. Therefore, it was used as the external measure to compare with the GBI.

Preliminary validity evidence. The mean number of effective adjustments correlated with posttest sum scores, $\rho = .23$, $p < .05$, suggesting players who scored higher on the posttest also tended to respond to in-game feedback in the desired direction. In addition, knowledge of force was positively related to mean force convergence ($\rho = .23$, $p < .05$) and negatively related to mean force deviation ($\rho = -.21$, $p < .05$), supporting the idea that these two indicators were measuring something meaningful (i.e., productive actions in the game related positively to knowledge, and unproductive actions related negatively to knowledge). Table 4 presents all pairwise correlations computed between posttest scores and each of the three GBIs.

Table 4

Spearman Correlations Between External Force Knowledge Measure and Game-Based Indicators

	1	2	3
1. Posttest sum scores	–		
2. Mean effective force adjustment	.23*	–	
3. Mean force deviation	-.21*	-.27**	–
4. Mean force convergence	.23*	.40***	-.44***

*p < .05 (two-tailed). **p < .01 (two-tailed). ***p < .001 (two-tailed).

Example 3: Bug-Inducing and Debugging Behaviors

Target knowledge or skill:	Debugging
Method used to derive the GBI:	Dissimilarity-based indicators
Theoretical justification:	Johnson et al.'s (1983) bug classification scheme defined four types of major programming bugs: (a) Missing is programming code that is required but omitted; (b) Redundant is programming code that is present but not required; (c) Misplaced is programming code that is required but in the wrong position; and (d) Malformed is programming code that is incorrect but in the correct position.
Validation approach:	Correlate GBIs to an external criterion measure
Summary of results:	• Knowledge of computational thinking concepts was positively correlated with productive debugging behavior (pretest: $\rho = .31$, $p < .01$; posttest: $\rho = .34$, $p < .01$) • Knowledge of computational thinking concepts was negatively correlated with unproductive debugging behavior (pretest: $\rho = -.23$, $p < .05$; posttest: $\rho = -.28$, $p < .05$).
Original study or source of data:	Feng, T., & Chung, G. K. W. K. (2022, to be presented 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]. American Educational Research Association (AERA) Annual Meeting, San Diego, CA, United States. Feng, T., & Chung, G. K. W. K. *Extracting debugging indicators based on distance to solution in a block-based programming game*. Manuscript under preparation.

Theoretical background. This example involves developing indicators of bug-inducing and debugging behaviors in a block-based programming game. We used Johnson et al.'s (1983) bug classification scheme to identify four types of major programming bugs—Missing, Redundant, Misplaced, and Malformed, where:

- Missing is programming code that is required but omitted.
- Redundant is programming code that is present but not required.
- Misplaced is programming code that is required but in the wrong position.
- Malformed is programming code that is incorrect but in the correct position; this often happens in more complex programming environments, such as the use of an incorrect data type during variable declarations (e.g., *int* versus *double* in C++).

GBI algorithm. In the programming game *codeSpark Academy*, players create programs by specifying the sequence of commands to achieve a goal. The game was designed to teach programming concepts to children aged between five and nine.

The commands are represented as blocks, where the player drags the block from the command palette to the code tray. Getting to an optional solution involves changing both the commands and the order of the block commands. To develop indicators of a potential debugging process, changes in commands were decomposed into (a) changes in code commands (or code content) and (b) changes in the structure (or ordering) of the code. Then player's code could be evaluated against a referent program (i.e., solution or gold standard) and a dissimilarity index computed.

Two indicators were developed to serve as inputs to more abstract indicators. A content indicator was created to represent the player's programming code in terms of commands present or absent. A structure indicator was created to represent the sequence commands.

Following Johnson et al.'s (1983) classifications, we first derived indicators of bug-inducing behaviors with two optimal solutions serving as the ideal references. The first two types of bugs—Missing and Redundant—are order-agnostic and focus on checking if the player has the right type and number of commands compared with the solutions (e.g., if a command in the solution is missing from the player's programming code). The remaining bug types, Misplaced and Malformed, focus on describing the order and form of the command sequence.

With the capability of detecting bug-inducing behaviors, we then developed secondary indicators targeting students' debugging processes as reflected by transitions between stages (stages 1 to 5) characterized by changes in bug-inducing behaviors. The five stages, from worst to best, are given in Table 5.

The indicator algorithm counted the number of forward transitions between categories (e.g., moving from Stage 1 to Stage 2), backward transitions (e.g., moving from Stage 2 to Stage 1), and self-transitions (e.g., moving from Stage 2 to Stage 2). We summed the credits to obtain player-specific game-based indicators that presumably reflected debugging.

Table 5

Stages of Programming Code Quality

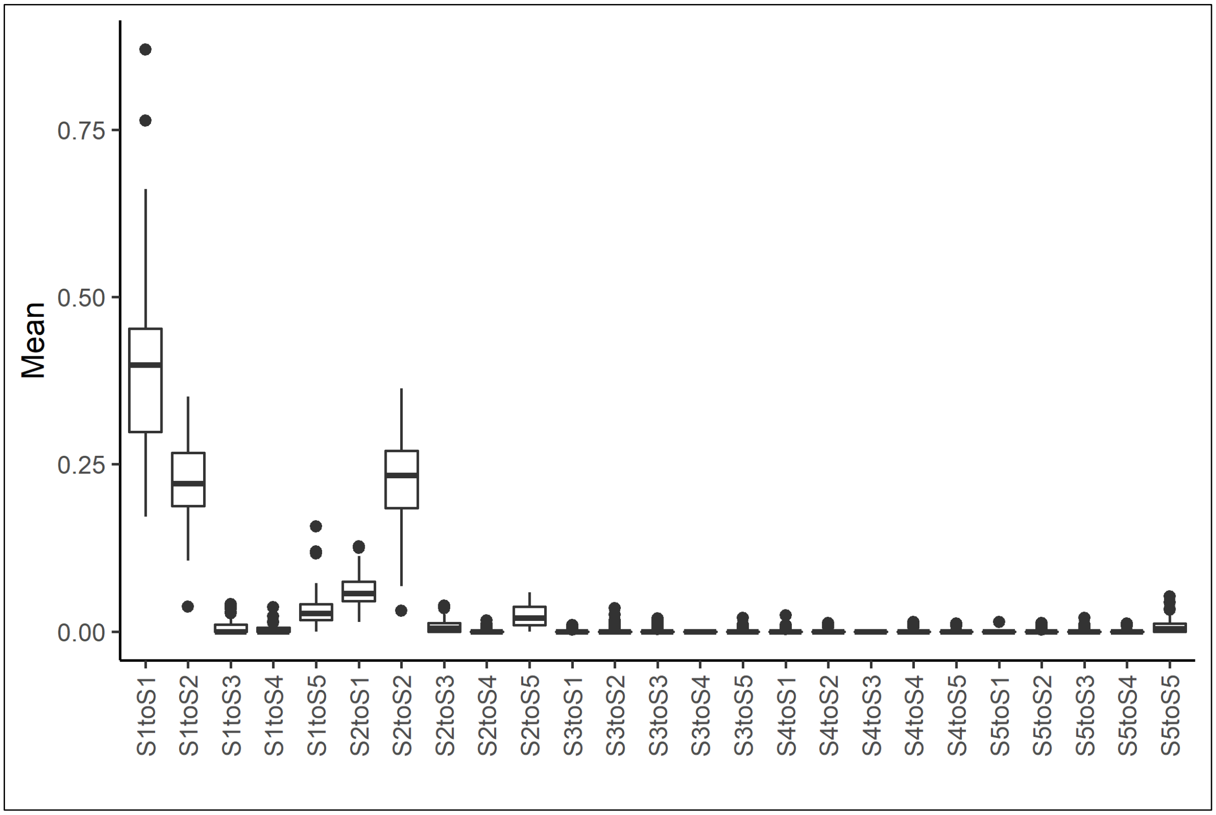
Stage	Description
1	Players' commands were at least 50% structurally dissimilar from the solution, implying the content was quite different.
2	There were missing commands, redundant commands, or both in players' programming code, and some commands were misplaced.
3	Players' commands were malformed.
4	Players' commands were only misplaced (no missing or redundant ones).
5	Players' commands were exactly the same—both in terms of content and structure—as the solution. The transition counts were averaged over the number of levels played for each player.

External criterion measure. The external criterion measure used was *TechCheck*, a 15-item multiple-choice assessment that measures computational thinking skills in domains that have been identified as developmentally appropriate for children between the ages of four and nine (Relkin et al., 2020). *TechCheck* had a Cronbach's alpha of .63 for the posttest.

Preliminary validity evidence. Figure 2 shows boxplots of averaged transition counts with self-transitions excluded. Table 6 presents the correlations between the external measures and game-based outcomes. Knowledge of computational thinking concepts was positively correlated with productive debugging behavior (represented by the mean number of times players moved forward) ($\rho = .34, p < .01$). Knowledge of computational thinking concepts was negatively correlated to unproductive debugging behavior (represented by the mean number of times players remained in the same stage) ($\rho = -.28, p < .05$).

Figure 2

Boxplots of Transitions Between Different Debugging Stages



Note. The x-axis labels denote the transition between two different stages. For example, *S1toS2* refers to the Stage 1 (S1) to Stage 2 (S2) transition.

Table 6

Spearman Correlations Between External Measures and Game-Based Outcomes

	1	2	3	4	5
1. Posttest sum scores	–				
2. Mean no. attempts	-.12	–			
3. Mean no. of correct first attempts	-.03	-.54***	–		
4. Mean no. in-game forward transitions	.34**	-.62***	.33**	–	
5. Mean no. in-game backward transitions	-.12	.06	.15	.19	–
6. Mean no. in-game self-transitions	-.28*	.55***	-.39***	-.95***	-.43***

*p < .05 (two-tailed). **p < .01 (two-tailed). ***p < .001 (two-tailed).

Example 4: Misconceptions

Target knowledge or skill:	Misconceptions about weight and pan balances
Method used to derive the GBI:	Game-specific indicators
Theoretical justification:	Metz (1993) identified and qualitatively described two misconceptions associated with young children's use of the pan balance: (a) displacing elements across pans and (b) the higher pan is the heavier object.
Validation approach:	Correlate GBIs to an external criterion measure
Summary of results:	• Knowledge of weight (posttest) was negatively related to the number of higher-is-heavier misconceptions (r = -.36, p < .01) • Learning of knowledge of weight (gain) was negatively related to the number of higher-is-heavier misconceptions (r = -.34, p < .05)
Original study or source of data:	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). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

Theoretical background. While many elementary school students can master the standard measurement tasks taught in school, few truly understand the measurement concepts (National Research Council., 2009). In areas ranging from physics to mathematics, children possess misconceptions that may hinder future learning (Perkins & Simmons, 1988). This example describes algorithms to detect misconceptions associated with using a pan balance. Metz (1993) qualitatively coded and defined general strategies in the development of weight-based problem solving for children ages 3 to 5. Two misconceptions identified by Metz were used in this study: (a) the higher pan is the heavier object and (b) displacing elements across pans.

GBI algorithm. For the games *Pan Balance*, we investigated two misconceptions: (a) heavier weights are those weights on the pan in the higher position, and (b) yanking or aligning one side of the balance with the other side to force equilibrium, which is how we operationalized displacing elements across pans. In the telemetry, we used information about the number of weights needed to beat the current round (target weight), the number of weights on the right side of the balance that children can interact with (scale weights), the number of weights on the table (table weights), and whether each placement or removal of one weight on or from the right side of the balance helps the player get closer to the target weight. From the telemetry, we computed the number of times a child placed a weight on the low pan or subtracted weight from the high pan. These in-game behaviors, in combination with other gameplay data, such as the number of times they received feedback about this error and total playtime, were used to identify the children with these misconceptions (either displacing elements across pans or the higher pan is the heavier object).

External criterion measure. The external measure used to compare our GBI indicators was a 20-item assessment using four items from the Child Math Assessment (CMA; Starkey et al., 2004) and three items from the KeyMath-3 Diagnostic Assessment (Connolly, 2007) and 13 items developed by our research group. The reliability of the assessment as measured by Cronbach's alpha was .82.

Preliminary validity evidence. To examine the relationship between the two misconception measures, correlations were computed between the misconception measures and the external measures of weight. Table 7 shows the intercorrelations between the GBIs and the external measure. The higher-is-heavier misconception was negatively related to the posttest knowledge of weight, $r(52) = -.36$, $p < .01$, and children's gain in weight knowledge, $r(52) = -.34$, $p < .05$. The significant negative correlation between the higher-is-heavier misconception and the knowledge gain is interesting because it suggests that children who were committing the misconception learned less from the game than other students.

Table 7

Intercorrelations Among Misconception Measures and External Measures of Weight

Measure	1	2	3	4	5
1. No. of higher-is-heavier misconceptions[a]	—				
2. No. of yank-or-align misconceptions[a]	.24	—			
3. Pretest[b]	-.04	-.06	—		
4. Posttest[b]	**-.36****	-.23	.58***	—	
5. Gain (Posttest – Pretest)[b]	**-.34***	-.19	-.30*	.59***	—
6. Time spent in *Pan Balance*[b]	.21	.08	.15	.18	.07

[a] $n = 53$. [b] $n = 66$.

*$p < .05$, two-tailed. **$p < .01$, two-tailed. ***$p < .001$, two-tailed.

Example 5: Deductive Reasoning

Target knowledge or skill:	Deductive reasoning
Method used to derive the GBI:	Game-specific indicators
Theoretical justification:	Device troubleshooting is a complex task that requires students to engage in reasoning strategies throughout the problem-solving process (Jonassen & Hung, 2006). Troubleshooting tasks require students to determine what information is needed for problem diagnoses and to reason with incomplete information—deductive reasoning.
Validation approach:	Correlate GBIs to an external criterion measure
Summary of results:	• Deductive reasoning was positively correlated with the ability to detect obstacles (sensitivity) ($\rho = .67$, $p < .001$) • Deductive reasoning was negatively correlated with players propensity to use lights (response bias) ($\rho = -.67$, $p < .001$). • Deductive reasoning was positively correlated with optimal decision-making (utility) ($\rho = .81$, $p < .001$).
Original study or source of data:	Chung, G. K. W. K., Redman, E. J. K. H., Eng, S., Feng, T., Michiuye, J. K., & Madni, A. (2019). *Developing innovative items to measure career readiness* (CRESST Report 861). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Theoretical background. Device troubleshooting is a complex task that requires students to engage in reasoning strategies throughout the problem-solving process (Jonassen & Hung, 2006). Troubleshooting tasks require students to determine what information is needed for problem diagnoses and to make judgments with incomplete information. Decision-making under uncertainty requires deductive reasoning.

In decision-making under uncertainty, students must choose an effective option from a mix of relevant and irrelevant options while observing the consequence of their choices. In a game task, the sources of uncertainty can be (a) exploration-driven: players have to explore unknown game areas; (b) solution-driven: players have to solve a problem or puzzle with unorganized or limited information; and (c) alignment-driven: players have to organize in-game objects or pieces of information in a spatial or conceptual configuration.

GBI algorithm. When the target construct requires decision-making under uncertainty, we can develop indicators that reflect a player's ability to make sound choices despite uncertainty or alternative noise. Signal Detection Theory (SDT) can be used to model the components of deductive reasoning. The most well-known application of SDT is in perceptual science, where SDT has been used to assess human performance in sensory detection and recognition tasks. SDT is also effective for measuring more complex scenarios, such as social threat perception that involves action classification and risk inference (Lynn & Barrett, 2014) and deductive reasoning (Trippas et al., 2014).

The game was designed for high school students and required them to program a (simulated) device controller to navigate a UFO from a starting point to a destination while avoiding obstacles and traps under cover of darkness. To help navigate the UFO, the player could use various controller tools (light, jump) to detect and avoid obstacles. The game increased in complexity, starting from no obstacles, to the inclusion of walls, to the inclusion of walls and vortexes.

To detect deductive reasoning, we used signal detection theory (SDT) within a utility framework to determine the GBIs. SDT indicators quantify a person's ability to discern information-carrying signals from noise, independent of a person's response bias (Macmillan & Creelman, 2005; Swets, Tanner, & Birdsall, 1961). The key indicators are sensitivity and response bias, both functions of hits and false alarms on decision-making tasks. SDT in a utility framework adds an additional variable, utility (Lynn & Barrett, 2014). Utility reflects optimal decision-making (vs. only accurate decisions) by incorporating the costs and benefits of decisions.

Sensitivity is the player's ability to discriminate between signal and noise. In the game where navigation was under cover of darkness, the signal was operationalized as obstacles that would cause a collision (i.e., a wall or vortex), the noise was operationalized as darkness, and the number of obstacles was unknown to the participants, which creates the perceptual uncertainty.

Response bias is the decision-making threshold influenced by the player's perceived uncertainty and the consequences of making a decision. In the game, response bias was operationalized as the player's propensity to use lights. The cost of being wrong results in the player needing to start over. Table 8 shows the classification table for signal and player perception, and Table 12 and Table 13 in the Appendix show the formulas for computing sensitivity, response bias, and utility.

Table 8

Classification Table for Sensitivity and Response Bias

	Player perception	
Signal	Signal perceived as absent	Signal perceived as present
Actual signal is absent	Correct rejection (CR): Perceiver believes there is no signal and the actual signal is absent.	False alarm (FA): Perceiver believes there is a signal, but the actual signal is absent.
Actual signal is present	Missed detection (MD): Perceiver believes there is no signal, but the actual signal is present.	Correct detection (CD): Perceiver believes there is a signal and the actual signal is present.

External criterion measure. Five items were adopted from Lawson's Classroom Test of Scientific Reasoning (Lawson, 2000). These items required deductive reasoning to solve the items correctly. All items were multiple-choice, and the total score for all items was used as the measure. Cronbach's alpha for the measure was .67.

Preliminary validity evidence. Table 9 shows the correlations among the external criterion measure and the three SDT indicators. The external measure of reasoning was highly correlated with the SDT indicators. Sensitivity, or the ability to detect obstacles, was positively correlated to reasoning, $\rho = .67$, $p < .001$. Response bias, or the propensity to use lights, was negatively correlated to reasoning, $\rho = -.67$, $p < .001$. Utility, which integrates costs and benefits with SDT indicators, resulted in a higher correlation with reasoning, $\rho = .81$, $p < .001$. This latter result is consistent with the operationalization of utility as optimal decision making.

Table 9

Correlations (Spearman) Among Task Performance and Processes Measures

Measure	1	2	3
1. External reasoning scale	–		
2. Sensitivity	.67**	–	
3. Criterion	-.67**	-.53*	–
4. Utility	.81***	.76***	-.76***

*p < .05, two-tailed. **p < .01, two-tailed. ***p < .001, two-tailed.

DISCUSSION

Our research question for this study was, to what extent do game-based indicators (GBI) relate to criterion measures of learning? We presented five examples where game-based indicators were developed based on a theory, and preliminary validity evidence was presented on the relationship between the GBI and an external criterion measure.

We found statistically significant correlations of varying magnitude depending on the study. The results are summarized in Table 10.

Table 10

Summary of Correlations Between GBI and External Criterion Measure by Example

Example	Type of GBI	Summary of Results
1. Speed and Accuracy	Common indicators	Performance indicators. Knowledge of numeracy was: • Positively related to performance in game (ρ =.4 to .6, p < .01 for correct first attempt at a solution) • Negatively related to the no. of failed solution attempts (ρ = -.4 to -.6, p < .01 for no. of incorrect solution attempts). Progress indicators. Knowledge of numeracy was: • Positively related to progress in the game (ρ = .67, p < .001) • Negatively related to the time taken to beat a level (ρ = -.57, p < .001).
2. Scientific Thinking	Dissimilarity-based indicators	Knowledge of force and motion was: • Positively related to effective force adjustments (ρ = .23, p < .05) • Positively related to productive game actions (ρ = .23, p < .05) • Negatively related to unproductive game actions (ρ = -.21, p < .05).
3. Debugging Behaviors	Dissimilarity-based indicators	Knowledge of computational thinking concepts was: • Positively correlated with productive debugging behavior (ρ = .34, p < .01) • Negatively correlated with unproductive debugging behavior (ρ = -.28, p < .05).
4. Mis-conceptions	Game-specific indicators	• Knowledge of weight (posttest) was negatively correlated with the no. of higher-is-heavier misconceptions (r = -.36, p < .01) • Learning (gain) was negatively correlated with the no. of higher-is-heavier misconceptions (r = -.34, p < .05)
5. Deductive Reasoning	Game-specific indicators	Deductive reasoning was: • Positively correlated with detection of obstacles (ρ = .67, p < .001) • Negatively correlated with propensity to use lights (ρ = -.67, p < .001). • Positively correlated with optimal decision-making (ρ = .81, p < .001).

Our results show several interesting findings. First, the direction of the correlations in all examples is in the expected direction. GBIs representing productive behavior are positively related to the external criterion measure, and GBIs representing unproductive behavior are negatively related to the external criterion measure. This finding is an important piece of validity evidence because it suggests that the GBIs measure (albeit weakly) the knowledge or skills measured by the external criterion measure.

The second interesting finding is that common indicators from Example 1 and those in Table 11 in the Appen appear to be sensitive to the criterion measure across a broad set of games (see Table 11). It may be that speed and time are fundamental variables underlying human performance. If so, then common indicators might serve a similar role as an effect size index: A standardized metric to compare games on their potential to serve as measures of knowledge or skill.

The third interesting finding is that we needed to refer to qualitative studies published nearly 30 years ago in Metz's case (1993) to find work on misconceptions around the pan balance that had a rich enough description of the actual child behavior and what the behavior meant in terms of learning. Similarly, we had to back nearly 40 years in Johnson's case (1983) to find a debugging scheme that was explicit enough about the rules and conditions to enable the classification of programming behavior into different bug categories. While purely speculative, we think there exists a trove of qualitative studies with detailed, comprehensive descriptions of student learning yoked to explicit student behavior under all sorts of conditions and outcomes (e.g., microgenetic analyses). Such studies are ideal for indicator and algorithm development, as illustrated by Example 4 (misconceptions), Example 3 (debugging), and to a lesser extent, Example 2 (scientific thinking), where we were able to translate detailed descriptions of the phenomena into algorithms.

Finally, we believe the most important finding is that we could develop GBIs (with moderate success) that have been defined by theory and not discovered through data-driven methods. We have demonstrated the operationalization of theory-based expected behaviors into algorithms by systematically transforming low-level clicks into successively more abstract and meaningful indicators. Furthermore, the finding that the indicators were significantly correlated with the external criterion measure at all is remarkable given how fine-grained the data are.

To elaborate on the significance of this effort: Making sense of telemetry and other fine-grained data—where one player can generate several hundred events during a single 60-minute session, and where the number of variables encoded in the telemetry can easily be in the hundreds—can be daunting. The data glut situation will only increase as more interactive systems become instrumented to collect telemetry. Some method is needed to impose structure on the data to enable the extraction of meaning. Further, the result should be explainable so that we can better understand precisely what a learner knows (or does not know) and can do (or cannot do). By having what we call traceable and inspectable indicators and algorithms, one can examine how an indicator is defined and operationalized, inspect what "ingredients" comprise inputs to algorithms and eventually statistical models, and evaluate analytically the extent to which the indicators represent the relevant, important, and educationally meaningful aspects of the construct being measured.

Limitations

There are many limitations to this work. First, the GBIs presented have all been from small-sample studies with sample sizes ranging from 20 to 150. The correlation coefficients reported may be unstable for the smaller samples. Second, most of the games were designed for young children, and thus there were a limited number of game mechanics. It is unclear how challenging it would be to develop theory-based GBIs for more complex learning games. Third, the games we have examined thus far were convenience samples. They were part of a current or past study and not a systematic sampling of games. The last limitation is the external criterion measure. The criterion measures used for the different examples sampled a broader range of knowledge than the GBI was targeting. This may partially explain the low correlations between the GBIs and criterion measures.

Next steps

Assuming theory-based GBIs are desirable and important, several next steps are needed. An immediate next step is to examine the alignment between the GBI algorithms and the items in the external criterion measures. This analysis may help explain the different magnitudes across the different examples. While the signs of the correlations were in the expected direction, the magnitude of the correlation coefficients varied widely across examples.

A second next step is to continue to develop GBIs for a broader range of learning games to get a better sense of the generalizability and utility of our approach. When does it not make sense to develop theory-based GBIs? Are theory-based GBIs better suited for some types of games (e.g., STEM games) than others?

A third next step is to examine GBIs to explicitly represent sequential processes. For example, GBIs that could represent the different phases of a problem-solving cycle would be extremely useful for two reasons: (a) to quantify a learner's skill at problem-solving; and (b) to characterize the degree to which a game or game level requires problem-solving skills, conditioned by player knowledge and experience variables.

Finally, a long-term next step is to accelerate the development GBIs. Algorithm development is labor-intensive, time-intensive, and highly specialized. Developing theory-based indicators requires expertise in coding, statistics, learning, measurement, and content. Indicator development involves extensive exploratory work. One way to accelerate development is to enable general researchers—who already understand the theory, learners, and the conditions that promote or detract from learning—to intuitively express expected behavioral patterns that indicate different learning processes and outcomes with the same degree of information as Metz (1993) had in her descriptions of preschoolers' behaviors and strategies.

The allure of using games for measurement purposes lies in its potential to offer rich forms of interaction and require learners to use domain knowledge and problem-solving strategies in an engaging way. Successful algorithm development requires understanding the underlying theory, how the theory is instantiated under different game conditions, and programming that can take into account these exigencies using fine-grained gameplay data.

REFERENCES

Ackerman, P. L. (1990). A correlational analysis of skill specificity: Learning, abilities, and individual differences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(5), 883–901.

Anderson, J. R. (1982). Acquisition of cognitive skills. *Psychological Review, 89*, 369–406.

Baker, E. L., Chung, G. K. W. K., & Delacruz, G. C. (2012). The best and future uses of assessment in games. In M. Mayrath, J. Clarke-Midura, D. H. Robinson, & G. Schraw (Eds.). *Technology-based assessments for 21st Century skills: Theoretical and practical implications from modern research* (pp. 229–248). Information Age Publishing.

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., & Parks, C. (2015). *Bundle 1 computational model analysis report* (Deliverable to PBS KIDS). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Chung, G. K. W. K., Redman, E. J. K. H., Eng, S., Feng, T., Michiuye, J. K., & Madni, A. (2019). *Developing innovative items to measure career readiness* (CRESST Report 861). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Connolly, A. J. (2007). *KeyMath-3 diagnostic assessment*. Pearson.

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

Ginsburg, H. P., & Baroody, A. J. (2003). *Test of early mathematics ability* (3rd ed.). Austin, TX: ProEd.

Gómez-Alonso, C., & Valls, A. (2008). A similarity measure for sequences of categorical data based on the ordering of common elements. In V. Torra & Y. Narukawa (Eds.), *Modeling decisions for artificial intelligence* (pp. 134–145). Springer.

Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science, 337*, 1623-1627.

Johnson, W., Soloway, E., Cutler, B., & Draper, S. (1983). *Bug catalogue: I* (Technical Report No. 286). Yale University, Department of Computer Science.

Jonassen, D. H., & Hung, W. (2006). Learning to troubleshoot: A new theory-based design architecture. *Educational Psychology Review, 19*, 77–114.

Lawson, A. E. (2000). Classroom Test of Scientific Reasoning [Multiple choice version] (Rev. ed.). Based on the development and validation of the classroom test of formal reasoning. *Journal of Research in Science Teaching, 15*, 11–24.

Lynn, S. K., & Barrett, L. F. (2014). “Utilizing” signal detection theory. *Psychological Science, 25,* 1663–1673.

Kato, P. M., & de Klerk, S. (2017). Serious games for assessment: Welcome to the jungle. *Journal of Applied Testing Technology, 18,* 1–6.

Kuhn, D. (2002). What is scientific reasoning and how does it develop? In G. Usha (Ed.), *Handbook of childhood cognitive development* (pp. 371–393). Blackwell.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Erlbaum.

McDermott, P. A., Green, L. F., Francis, J. M., & Stott, D. H. (2000). *Preschool Learning Behaviors Scale*. Edumetric and Clinical Science.

McDermott, P. A., Leigh, N. M., & Perry, M. A. (2002). Development and validation of the Preschool Learning Behaviors Scale. *Psychology in the Schools, 39*, 353–365.

Metz, K. E. (1993). Preschoolers' developing knowledge of the pan balance: From new representation to transformed problem solving. *Cognition and Instruction*, *11*, 31–93.

Mislevy, R. J., Corrigan, S., Oranje, A., DiCerbo, K., & Bauer, M. I. (2014). *Psychometrics and Game-Based Assessment*. 26.

National Research Council. (2009). *Mathematics learning in early childhood: Paths toward excellence and equity.* Washington, DC: The National Academies Press. https://doi.org/10.17226/12519.

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

PBS KIDS. (2020). *PBS KIDS Social & Emotional Learning (SEL) and Character Development Frameworks* (v2.0).

Perkins, D. N., & Simmons, R. (1988). Patterns of misunderstanding: An integrative model for science, math, and programming. *Review of Educational Research, 58*, 303-326.

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). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

Redman, E. J. K. H., Schenke, K., Chung, G. K. W. K., Parks, C. B., Michiuye, J. K., Feng, T., Chang, S. M., & Cai, L. (2019). *Analytics validation final report* (Deliverable to PBS KIDS). UCLA/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*. UCLA/CRESST.

Relkin, E., de Ruiter, L., & Bers, M. U. (2020). *Techcheck*: Development and validation of an unplugged assessment of computational thinking in early childhood education. *Journal of Science Education and Technology, 29*, 482-498. https://doi.org/10.1007/s10956-020-09831-x

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.

Sao Pedro, M. A., de Baker, R. S. J., Gobert, J. D., Montalvo, O., & Nakama, A. (2013). Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. *User Modeling and User-Adapted Interaction*, *23*(1), 1–39. https://doi.org/10.1007/s11257-011-9101-0

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

Simon. H. (1977). *Models of discovery*. Reidel.

Starkey, P., Klein, A., & Wakeley, A. (2004). Enhancing young children's mathematical knowledge through a pre-kindergarten mathematics intervention. *Early Childhood Research Quarterly, 19*, 99–120.

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68*, 301–340.

Trippas, D., Handley, S. J., & Verde, M. F. (2014). Fluency and belief bias in deductive reasoning: new indices for old effects. *Frontiers in Psychology, 5*, 1–7.

van der Graaf, J., Segers, E., & Verhoeven, L. (2018). Individual differences in the development of scientific thinking in kindergarten. *Learning and Instruction, 56,* 1–9.

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review, 27*, 172–223.

APPENDIX

Table 11

Correlations Between Progress Measures and Performance Measures and External Outcome Measures

Game	Progress measures		Performance measures		
	No. of levels completed	Mean time spent on a level	No. of correct first attempts	No. of correct attempts overall	No. of incorrect attempts overall
PBS KIDS Cat in the Hat Builds That					
Perform task as outcome					
Bridge-a-rama	.16	-.01	.16	.15	.12
Slidea-ma-zoo	.24*	-.32**	.32**	.18	.11
Sorta-ma-gogo	.18	-.27*	unavailable	.22	-.05
Knowledge task as outcome					
Bridge-a-rama	.20*	-.03	.26**	.21*	-.10
Slidea-ma-zoo	.13	-.29**	.16	.12	-.15
Sorta-ma-gogo	.05	-.12	unavailable	.11	-.18
PBS KIDS Cat in the Hat Knows a Lot About That! [a]					
Perform task as outcome					
Bridge-a-rama	.10	.11	.03	.10	.08
Slidea-ma-zoo	.08	.04	.05	.09	.07
Sorta-ma-gogo	.04	.01	unavailable	.04	.05
Lens on Science task as outcome					
Bridge-a-rama	.08	.00	-.01	.08	-.22**
Slidea-ma-zoo	.29***	.04	.28***	.29***	.02
Sorta-ma-gogo	.17*	-.31***	unavailable	.18*	-.14
PBS KIDS Ruff Ruffman[b]					
Fish Force	.36***	-.28**	.54***	.48***	.18
PBS KIDS Connected Learning					
Air Show	.39**	.23	unavailable	.27*	.14
All Star Sorting	-.13	-.20	unavailable	-.12	-.24
Bird Measurer	.48***	.08	unavailable	.47***	.19
Bubble Bath	.40**	.22	unavailable	.40***	.30*
Cart Balancer	.45***	.19	unavailable	.45***	.26*
Cauldron Filler	.44***	—	unavailable	.39**	.24

	Progress measures		Performance measures		
Game	No. of levels completed	Mean time spent on a level	No. of correct first attempts	No. of correct attempts overall	No. of incorrect attempts overall
Chest Sorter	.51***	.46***	unavailable	.49***	.47***
Chow Time	.42***	-.04	unavailable	.38**	.17
Crystals Rule	.17	-.02	unavailable	.18	-.01
Dino Dive	.38**	.07	unavailable	.38**	.23
Happy Camel	.30*	.07	unavailable	.29*	.12
Leaf Leader	.43***	.07	unavailable	.38**	.26*
Mushroom Sorter	.21	.02	unavailable	.24	-.20
Pan Balance	.49***	.10	unavailable	.44***	.20
Scrub A Dub	.33**	-.11	unavailable	.30*	.19
PBS KIDS Curious George [c, d]					
Blast Off	.24	-.57***	.42**	.11	-.40**
Apple Picking	.23	-.28*	.39**	.16	-.50***
Meatball Launcher	.67***	-.05	.63***	.67***	-.56***
Math Games (middle school) [d]					
Save Patch	.44**	-.37*	.55**	.43**	-.45**
Tlaloc's Book	.27**	-.29**	.31**	.28**	-.10**
AlgebRock	.50**	-.49**	.54**	.50**	-.32**
Math and physics game and simulation prototypes (adults) [d]					
Exponents	unavailable	unavailable	.59**	.40*	-.39*
LearnForm	.21	unavailable	.44**	.23	-.09

[a] Grindal et al. (2019) (for performance task: sum scores; for Lens: posttest scores). [b] Redman et al. (in press). [c] Chung and Parks (2015a, 2015b). [d] Chung and Roberts (2018).
*$p < .05$ (two-tailed). **$p < .01$ (two-tailed). ***$p < .001$ (two-tailed).

Table 12

Computation and Interpretation for Sensitivity and Criterion (Macmillan & Creelman, 2005)

Measure	Computation	Interpretation
Sensitivity (*d*' or *d*-prime)	The standardized difference between the means of the correct detection and false alarm rates, where CD and FA are expressed as z-scores: $d' = z(CD) - z(FA)$	*d*' = 0 indicates the perceiver cannot discern the signal from noises. The higher *d*', the higher sensitivity. In the context of the UFO game, *d*' reflects the difference between detecting an obstacle. Higher values of *d*' indicate the player can more often detect obstacles.
Response Bias	$c = -\frac{z(CD) + z(FA)}{2}$	*c* = 0 indicates the correct detection and false alarm rates are equal. A negative value for *c* indicates that the false alarm rate exceeds the correct detection rate, and a positive value for *c* indicates the correct detection rate exceeds the false alarm rate. A higher *c* indicates a more conservative approach. In the context of the UFO game, *c* is the decision threshold above which the player will conclude there is an obstacle and thus use the light function.

Table 13

Computation of the Expected Value Estimator. Reproduced from Lynn, Wormwood, Barrett, and Quigley (2015, pp. 6)

Measure	Computation	Interpretation
Expected value, *U*, of a decision criterion at signal value x_i.	$U(x_i) = \alpha hp[\text{CD}] + \alpha mp[\text{MD}] + (1 - \alpha)ap[\text{FA}] + (1 - \alpha)\, jp[\text{CR}]$	Optimal or near-optimal decision making.

Note. Base rate parameters: α = base rate or relative probability of encountering a signal, 1 – α = relative probability of encountering a signal from the noise. Payoff parameters: *h* = benefit of correct detection, *m* = cost of missed detection, *a* = cost of false alarm, *j* = benefit of correct rejection. Similarity parameters: p[CD] = correct detection rate, p[MD] = missed detection rate, calculated by 1 – p[CD], p[FA] = false alarm rate, p[CR] = correct rejection rate, calculated by 1 – p[FA].