JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

DOI: 10.1111/jedm.12396

Journal of Educational Measurement
Spring 2026, Vol. 63, No. 1, e12396

Sensemaking of Process Data from Evaluation Studies of Educational Games: An Application of Cross-Classified Item Response Theory Modeling

Tianying Feng and **Li Cai**
UCLA/CRESST, Los Angeles, CA, United States

Process information collected from educational games can illuminate how students approach interactive tasks, complementing assessment outcomes routinely examined in evaluation studies. However, the two sources of information are historically analyzed and interpreted separately, and diagnostic process information is often underused. To tackle these issues, we present a new application of cross-classified item response theory modeling, using indicators of knowledge misconceptions and item-level assessment data collected from a multisite game-based randomized controlled trial. This application addresses (a) the joint modeling of students' pretest and posttest item responses and game-based processes described by indicators of misconceptions; (b) integration of gameplay information when gauging the intervention effect of an educational game; (c) relationships among game-based misconception, pretest initial status, and pre-to-post change; and (d) nesting of students within schools, a common aspect in multisite research. We also demonstrate how to structure the data and set up the model to enable our proposed application, and how our application compares to three other approaches to analyzing gameplay and assessment data. Lastly, we note the implications for future evaluation studies and for using analytic results to inform learning and instruction.

Sensemaking of Data From Game-Based Evaluation Studies

Understanding data from evaluation studies of educational games *to answer questions about effectiveness and improvement* involves, if not necessitates, sensemaking. Sensemaking is a deliberate effort to construct a plausible understanding of differences, complex situations, or ill-structured problems (Dervin, 2003; Klein et al., 2007; Pirolli & Russell, 2011; Weick et al., 2005). Sensemaking starts with posing a frame. To frame is to initiate with a story, perspective, or framework that guides how one explores, defines, connects, and interprets data (Klein et al., 2007). One goal of sensemaking is to inform practices (Dervin, 2003) and actions (Klein et al., 2007; Weick et al., 2005). We use sensemaking as the conceptual underpinning of our paper and as a guide for prioritizing the preconditions upon which our paper is built.

Sensemaking occurs when we grapple with information embedded in gameplay process data. Event-based gameplay process data track individuals' moment-to-moment choices, interactions with in-game elements, and other game-related information, including the timing of each event and details about the elements interacted with (Chung, 2015). The resulting data can contain hundreds of rows

associated with each player attempting one in-game task or puzzle. Well-instrumented (gameplay) process data contain information—construct-relevant patterns and processes—that we can leverage to help infer how different individuals approach a task. The challenge lies in how we extract, analyze, and interpret such information to draw valid inferences about what individuals know or learn (Greiff et al., 2015; Lindner & Greiff, 2023).

Sensemaking also occurs when we integrate gameplay *process* information with other *outcome* (Bergner & Davier, 2019) or *product* (Levy, 2020; Zumbo et al., 2023) information obtained from mediums like traditional assessments. This integration may serve several functions, such as holistically gauging the instructional effectiveness of a game-based intervention or its absence thereof, exploring the relationship between the process and the outcome, and of equal importance, answering the question: "How and why did learning, growth, or change (not) occur?" Statistical modeling is one tool that accomplishes the integration. What is more, integration aimed at sensemaking demands an understanding of how factors beyond data analysis and model construction interact. Some of these factors are not exclusive to game-based research, including (a) alignment in content and cognitive demand features between the game and external assessments (Baker et al., 2008; Mislevy et al., 2015); (b) game or task design features (Mislevy et al., 2014, 2015; Plass et al., 2015), user-interface design features (Chung & Baker, 2003), and their effects on gameplay and cognition; (c) data instrumentation (Bergner & Davier, 2019; Chung, 2015); and (d) use of theory-informed or construct-sensitive process information (Goldhammer et al., 2021; Lindner & Greiff, 2023). Overlooking these factors, such as poor data instrumentation and alignment, compromises data quality and validity of inferences made through a model that aims to integrate gameplay and assessment information for sensemaking.

Of the many factors, we prioritize three as the preconditions for fruitful sensemaking and for the cross-classified item response theory (IRT) modeling approach advocated in this paper. These preconditions are: (a) alignment in content and cognitive demand features between the game and the assessment, (b) use of diagnostic or theory-informed process information, and (c) provision of a unified and flexible modeling framework. All three preconditions are crucial for creating a coherent analytical framework that enables us to use gameplay as a diagnostic tool, derive interpretable results, and provide feedback to inform game design, student learning, and instruction. In what follows, we discuss the existing shortfalls in meeting one or more of these preconditions.

Existing Shortfalls in Meeting Three Preconditions for Data Sensemaking

Shortfall 1: Lack of Alignment between Game Design and Assessment Design

The most critical shortfall that impedes sensemaking is the lack of alignment between what is measured and instructed by the game-based intervention, and what is assessed by the assessment. The design of the learning system (e.g., educational games) is often distinct from that of the assessment system. This observation is articulated directly (Arieli-Attali et al., 2019) or through calls, recommendations, and needs (e.g., Darling-Hammond et al., 2013; Gane et al., 2018; Foster & Piacentini,

2023; National Research Council, 2001; Pellegrino & Quellmalz, 2010). This disconnection reduces the probability of leveraging rich diagnostic information gained from process data to provide feedback to the intended users, such as students and instructors. It can lead to unnecessarily burdensome summative assessments when information about progress and achievement is already available from process data, a perspective akin to what Pellegrino and Quellmalz (2010) argued for technology-enabled assessments. The lack of alignment in the content and cognitive demand features between the game and the assessment, as well as between the in-game tasks and the assessment items, can also limit the range of questions and analyses available for exploration.

Shortfall 2: Underuse of Diagnostic or Theory-Informed Process Information

The second shortfall stems from the first. By "underuse," we mean a tendency to prioritize the analysis of data from traditional assessments as the primary or exclusive source of evidence, whether in game-based research (Garcia et al., 2020; Petri & Gresse von Wangenheim, 2017) or more broadly in studies with access to both traditional assessment and process data (e.g., computer-based assessments; Greiff et al., 2015). Little attention is given to understanding learners' experiences within the digital medium, and in the case of game-based evaluation research, within the intervention itself. This neglect renders conceptually meaningful information embedded in process data an "often-noted but seldom used potential" (Greiff et al., 2015, p. 93).

By "underuse," we also mean that the use of diagnostic process information remains limited when compared to information on time, response accuracy, or generally, timing and counts of low-level events (Greiff et al., 2015). Time spent on a task and response correctness are jointly modeled in cognitive diagnosis and psychometric modeling research (De Boeck & Jeon, 2019; Ercikan et al., 2020; Jiao et al., 2019; Lee & Jia, 2014; van der Linden, 2007). Indicators of time and event counts are also frequently featured in substantive research (e.g., Blanié et al., 2020; Chen et al., 2020; Cagiltay et al., 2015; Gauthier et al., 2015; Goldhammer et al., 2014; Hahnel et al., 2023; Hautala et al., 2020; Kiili et al., 2018; Shute & Rahimi, 2021; Tenorio Delgado et al., 2016).

In comparison, diagnostic or theory-informed process information is underreported and underused. This kind of information includes indicators of (mis)conceptions (Chung & Feng, 2023; Kerr, 2014), strategies (Chung & Baker, 2003; Greiff et al., 2015; Wüstenberg et al., 2012), expert-defined rules (Hao et al., 2015), or cognitively meaningful patterns (Liu & Israel, 2022). The requirements for substantive knowledge and technical expertise affect how feasible it is to translate theories into gameplay and to extract diagnostic process information. Research on computer-based assessments has also articulated similar challenges (Greiff et al., 2015; Lindner & Greiff, 2023).

Shortfall 3.1: Lack of a Unified Modeling Framework

We divide the third shortfall into two parts, Shortfalls 3.1 and 3.2, both of which concern modeling. Gameplay and assessment data are often not jointly analyzed

to connect the assessment and the intervention. This also means that the estimation of the intervention effect is based on only a subset of the collected information or separate analyses, the findings of which are not integrated or cannot be integrated.

Analyses used in game-based research tend to employ a two-stage procedure to examine the relationships among indicators derived from gameplay process data and assessment scores. In stage one, indicators are aggregated to the person level. In stage two, aggregated indicators are correlated with assessment scores via correlational analysis (Petri & Gresse von Wangenheim, 2017; Zhu et al., 2023), included as covariates in regression analysis (Hautala et al., 2020; Weiner & Sanchez, 2020), or used in a predictive framework with deep learning models (Min et al., 2020). Notable exceptions to the aforementioned analyses include (a) an analysis by Kerr and Chung (2012b), which investigated how individuals' overall game-based performance scores mediated the relationship between pretest and posttest sum scores; (b) an analysis by Reese et al. (2015), which used multilevel modeling to examine the velocity and acceleration in individuals' progresses toward the game goal; and (c) an analysis by Levy (2019), which applied dynamic Bayesian network modeling to nonaggregated, longitudinal game-based indicator data.

It is also worth noting that outside the game-based research context, more modern statistical and computational approaches have been developed to make use of process information (Jiao et al., 2021; Lindner & Greiff, 2023). For instance, in computerized assessments, process information has been used or integrated to refine assessment information and improve test reliability (Tang et al., 2020; Xiao et al., 2022; Zhang et al., 2023). Techniques used in categorical sequence analysis (Abbot & Tsay, 2000) are used to extract latent features from categorical process data. Along the line of embedding and dimensionality reduction, latent space modeling has been applied to process data (Chen et al., 2022).

Shortfall 3.2: Lack of a Flexible Modeling Framework

Commonly used statistical tools in game-based research, such as correlational analysis and multiple regression modeling, often fail to address data complexities stemming from features of study design. Nor do they provide the flexibility for researchers to explore additional hypotheses. In multisite evaluation research, four types of dependency can occur in the collected assessment data (Cai et al., 2016), apart from complexities that arise in gameplay data. First, there is a dependency between the underlying constructs being measured over time. Second, when the same set of items is administered to students repeatedly, there is item-level residual dependence. Third, the assumption of full exchangeability of individuals across experimental conditions often fails to hold, particularly in randomized controlled trials (RCTs) of learning games, where varied degrees of student learning are expected to occur as a result of the intervention. Fourth, there is a dependency among individuals nested in sites. In addition to the acknowledged dependencies, researchers may want to investigate how characteristics at the site, in-game task, or person level influence individuals' performance and progress.

This Paper

We pose one frame for sensemaking of data collected in evaluation studies of educational games. This frame presupposes the first two preconditions are met: (a) alignment (Chung et al., 2014; Center for Advanced Technology in Schools, 2012; Vendlinski et al., 2010) and (b) use of diagnostic process information (Kerr, 2014; Kerr & Chung, 2012a). With this frame, we address the third precondition about modeling.

Our frame consists of two parts. First, among various methods for extracting information from process data, we use indicators of knowledge misconceptions derived from gameplay processes (Kerr, 2014). We use these indicators as a means to "notice and bracket" (Chia, 2000) diagnostic insights from gameplay, moving away from simpler indicators that often fall short in making "students' behaviours and thought processes visible" (Foster & Piacentini, 2023, p. 37). Second, we present a new application of cross-classified IRT modeling to integrate data from multiple sources, specifically data of gameplay and assessments.

The new application addresses issues and data complexities mentioned above. It jointly models individuals' responses on game-based diagnostic indicators and responses to assessment items, and incorporates gameplay information when gauging the intervention effect of an educational game. It also relates individuals' game-based performance to changes in assessment outcomes and accounts for the nesting of individuals within sites in multisite studies.

In the following sections, we present the notations and the data structure that combines gameplay indicator data and assessment item data. We introduce the general framework and components of cross-classified IRT modeling (Huang & Cai, 2024; van den Noortgate et al., 2003). We then compare our proposed application with three other approaches applied to a data set collected from a multisite RCT of math games (Chung et al., 2014) to highlight advantages of our application. Lastly, we note the implications for future (game-based) evaluation studies and for using analytic results to inform learning and instruction.

Data Structure: Game-Based Indicator Data Combined with Item-Level Assessment Data

Game-Based Indicator Data

What are game-based indicators? Game-based indicators are *observable variables* (Mislevy et al., 2014) derived from gameplay process data. These indicators capture different aspects of players' interactions with a game. Examples include indicators of time spent on tasks, low-level event counts, patterns, strategies, and knowledge misconceptions. Often, we derive one or more game-based indicators for each in-game task and for each player.

What are in-game tasks? An in-game task is a game level that is explicitly defined during the game's design and creation process. We use *in-game tasks*, as opposed to *game levels*, to distinguish levels in a game from levels in the multilevel modeling framework. We use *tasks* to also emphasize that researchers can adopt our

Table 1
Example of Indicator Data with One Binary Indicator

Player	Task 1	Task 2	Task 3
1	1	0	1
2	0	1	0
3	1	1	1

Table 2
Example of Indicator Data with Three Binary Indicators (Semi-Long Format)

Player	Task	Indicator 1	Indicator 2
1	1	1	0
1	2	1	0
1	3	0	1
...			
3	1	0	1
3	2	1	1
3	3	1	0

proposed application not only to analyze game-based indicators but also data from other mediums such as simulations.

Data structure with one indicator. The simplest game-based indicator data consists of one binary game-based indicator derived per in-game task and per player. Let us consider a scenario where three players have completed three in-game tasks and provided responses on one binary indicator. The indicator shows whether a player used a particular strategy to complete a task. Table 1 presents the indicator data associated with this example.

In this example, each row contains one player's data, and each column corresponds to one in-game task. A value of 1 indicates that the player used the strategy of interest to complete the task, while a value of 0 indicates that the player did not use the strategy.

Data structure with more than one indicator. Often, we derive more than one indicator to describe different aspects of individuals' gameplay. Suppose that for the same three tasks (Tasks 1-3) and the same three players (Players 1-3), we derive multiple indicators to measure different aspects of player performance. Without loss of generality, we assume that there are two binary indicators that capture the completion (1) or failure (0) of two specific objectives of a task. Table 2 shows the data in a semi-long format, where each row corresponds to a unique player-task combination and contains the player's responses on two game-based indicators.

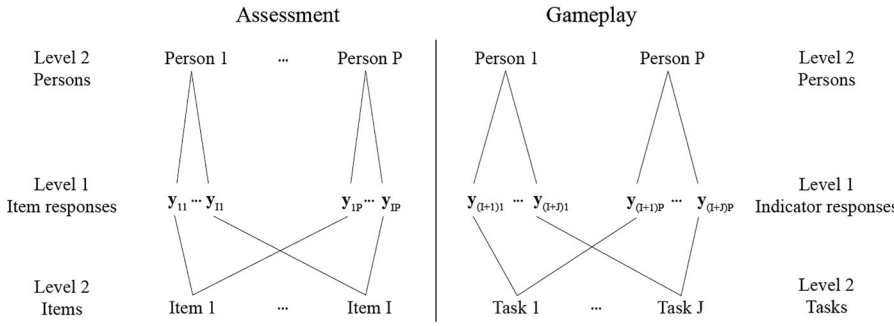

Figure 1. Cross-classified structure of assessment item data and gameplay indicator data.

Game-Based Indicator Data as Cross-Classified Data

Individuals' in-game performance is affected by their knowledge and skills and the characteristics of the tasks they encounter. This observation is also evident in the data structure demonstrated in the earlier examples, where the responses on indicators are simultaneously organized or classified by both players (persons) and tasks. The relationship between persons and tasks does not exhibit a clear hierarchical structure. Instead, persons and tasks represent two distinct sources of variation. Therefore, it is appropriate to consider the responses on the derived indicators as cross-classified, or simultaneously influenced, by both persons and tasks.

Figure 1 shows game-based indicator data and item response data as two examples of cross-classified data. Responses on game-based indicators are cross-classified by persons and in-game tasks, while responses to the assessment items are cross-classified by persons and items. Moreover, if we represent an individual's assessment-based competency and game-based performance as two latent variables, we can correlate the two variables to connect in-game performance with assessment outcomes, and vice versa.

Modeling in-game tasks as a random component offers several advantages, similar to modeling test items as random (e.g., De Boeck, 2008; De Boeck & Wilson, 2004; van den Noortgate et al., 2003). First, the random task approach is useful when a vast pool of in-game tasks are generated by manipulating specific design elements, and the tasks of interest are a sample drawn from this pool. Second, modeling tasks as random instead of fixed reduces the computational burden by reducing the number of parameters to be estimated, while still considering the task effects on gameplay. Consequently, the random task approach may be more viable for studies with limited sample sizes. When adopting the random task approach, however, we are not interested in the individual tasks themselves but rather in describing their variability. Lastly, by regressing the random task variable on design variables, researchers can explore the extent to which different design variables explain the task effects.

Structure of the Combined Data

We now combine the game-based indicator data with item-level data from traditional assessments. Table 3 shows a general representation of the combined data in

Table 3
Structure of the Combined Data (Wide Format)

		Assessment Block Item $1 \cdots i \cdots I$	Gameplay Block Task $1 \cdots j \cdots J$
Person	1 ⋮ p ⋮ P	$\mathbf{y}_{ip}$	$\mathbf{y}_{jp}$

the wide format. In this table, two blocks of data are presented: the assessment block of items and the gameplay block of in-game tasks. Here, P is the number of persons, I is the number of items, and J is the number of in-game tasks. The boldface $\mathbf{y}_{jp}$ indicates that one or more indicators can be derived for each Task j. For example, we may be interested in whether a person uses Skill A and Skill B to complete an in-game task, creating one indicator per skill. Similarly, the boldface $\mathbf{y}_{ip}$ indicates that one or more responses can be made for each Item i. One example is a test-retest setting, where individuals respond to a set of identical items at two different time points (e.g., pretest and posttest).

Example of the Combined Data

To bridge the previous discussions with a concrete example, let us consider a data set containing three individuals ($P = 3$) who attempted two pre- and post-assessment items ($I = 2$) and two in-game tasks ($J = 2$). For each of the tasks, we have also derived two indicators ($K = 2$) using the gameplay process data. All item and indicator responses are dichotomously scored.

To connect this example with the data set discussed in the "Empirical Data Analysis" section, where we present the new application of cross-classified IRT modeling, we assume that the individuals (Persons 1-3) belong to different schools, where individuals within the same school share more similar experiences compared to individuals from different schools. Specifically, Person 1 and Person 2 are from School 1, and Person 3 from School 2. Table 4 presents the data in the wide format. Table 5 presents the data in the semi-long format. For subsequent modeling and analyses, we use data in the semi-long format.

In Table 5, we use *Variables* to denote the columns of observed responses on game-based indicators or assessment items. It is important to note that *Variables* (capitalized) refers specifically to observed responses (or observed variables) and should not be confused with any discussion of latent variables. We use the term *Block* to denote which block of data an observed variable belongs to, distinguishing between responses associated with items and with in-game tasks. The *Elements* within each block are the items or tasks, labeled using the same suffixes shown in Table 4. For example, the first item in the assessment block is denoted as I1.

Table 4
Example Combined Data (Wide Format)

		Assessment (Block 1)				Gameplay (Block 2)			
School	Person	Pre_I1	Pre_I2	Pst_I1	Pst_I2	J1_K1	J1_K2	J2_K1	J2_K2
1	1	0	0	1	0	1	0	0	0
1	2	1	0	1	0	0	0	1	0
2	3	0	1	1	1	0	1	0	0

Note. Pre: pretest; Pst: posttest; J: in-game task. The suffix K denotes the game-based indicators, and the suffix I denotes the assessment items.

Table 5
Example Combined Data (Semi-Long Format)

			Variables					
			Item Responses		Indicator Responses		School Dummy Variables	
Person	Block	Element	Pre	Pst	K1	K2	Sch1	Sch2
1	1	I1	0	1			1	0
1	1	I2	0	0			1	0
1	2	J1			1	0	1	0
1	2	J2			0	0	1	0
2	1	I1	1	1			1	0
2	1	I2	0	0			1	0
2	2	J1			0	0	1	0
2	2	J2			1	0	1	0
3	1	I1	0	1			0	1
3	1	I2	1	1			0	1
3	2	J1			0	1	0	1
3	2	J2			0	0	0	1

Pre: pretest; Pst: posttest; Sch: school; K: game-based indicator.

Lastly, we add school dummy variables (Sch1 and Sch2) to indicate the respective school that each person belongs to. For example, Person 1 is from School 1, given that every cell associated with Person 1 under the Sch1 column has a value of one. This person has responses on two binary indicators (K1 and K2) from two in-game tasks (J1 and J2) and responses to two items (I1 and I2) administered during both pretest and posttest.

A General Modeling Framework Using Cross-Classified Item Response Theory Models

Overview

To jointly analyze data from different sources, such as game-based indicator data and item-level assessment data, we consider three sources of influence on

individuals' observed responses: (a) the blocks, such as one block of items and another block of in-game tasks; (b) the persons; and (c) the sites. Persons are nested in sites, such as students in different classrooms or schools.

To account for these sources of influence in modeling, we model the blocks and the persons as latent variables, and we include school dummy variables to account for the nesting of persons in schools. When examining how students respond to assessment items and in-game tasks, we assume that the person-specific latent variables reflect the ability, knowledge, or skills of the students, and the block-specific latent variables reflect the relative difficulties of in-game tasks and assessment items.

Here, we have a cross-classified structure: the observed responses are cross-classified by blocks and persons. The observed responses are situated on Level 1, while persons and blocks are situated on Level 2. We incorporate school dummy variables to address the nesting of persons within sites and estimate a separate intercept for each school, as with a fixed effects model. This approach differs from treating school effects as random, where the emphasis is on summarizing the distribution of school effects using a group mean and variance.

Notations

Let s index sites ($s = 1, \ldots, S$), p index persons ($p = 1, \ldots, P$), b index blocks ($b = 1, \ldots, B$), e_b index elements in Block b (e.g., an item in the assessment block; $e_b = 1, \ldots, E_b$), and v index the vth Variable (e.g., one of the Variable columns in Table 5). With the data structure demonstrated in Table 5, each element is essentially associated with only one block—either the gameplay or the assessment block. Therefore, we can drop the Block subscript b, and let y_{veps} denote the response on Variable v for Element e by Person p in Site s. The person-specific latent variables are referred to as the *cluster-side* latent variables (e.g., η_p). The item- and task-specific latent variables are referred to as the *block-side* latent variables (e.g., ξ_e).

The Linear Predictor

We denote the conditional probability of a response given relevant parameters as π, omitting all subscripts. In the context of modeling binary responses, we consider an IRT model with the following basic form:

$$\pi = \frac{1}{1 + \exp(-z)}. \tag{1}$$

In Equation 1, z is the linear predictor. We will gradually build the desired cross-classified IRT model, beginning with the simplest form of the linear predictor and moving toward the full model specification.

The simplest form. We start by considering a simple scenario where there is no nesting of individuals within sites, and we do not include any explanatory predictors in the latent regression equations. We also assume a one-parameter logistic (1PL) IRT model with one person-specific latent variable and one block-specific latent variable and the corresponding loadings, or slopes, are constrained to be 1. We use z_{vep} to denote the linear predictor in a cross-classified 1PL model for Variable v of Element e (in Block b) responded by Person p. The block subscript b is omitted

when each element is only associated with one block. With the simplifications, z_{vep} can be written as:

$$z_{vep} = \alpha_v + \xi_e + \eta_p. \tag{2}$$

Two sources of variation are present: one arising from the persons and the other arising from the blocks. The person-specific latent variable is $\eta_p \sim N(0, \sigma^2_{Person})$, which varies over the persons. The block-specific latent variable is $\xi_e \sim N(0, \sigma^2_{Block})$, which varies over elements in a given block. Again, the block-specific subscript b is omitted, given that all elements belong to only one block. Although we will not consider this in detail, the modeling framework could accommodate a generalizability theory (G-theory) like interaction term between persons and blocks. It is crucial, however, to have enough replications, determined by design, within each of the cross-classified cells to help disentangle this interaction from other within-cell unidentified sources of variation ("error"); without replications, they would be confounded (e.g., Shavelson & Webb, 1991, pp. 20-23; Raudenbush & Bryk, 2002, pp. 377-378).

Substantively speaking, $\hat{\eta}_p$ is the estimated score of Person p's latent factor of interest. This latent factor could represent an individual's underlying ability, skill, or knowledge. For a block of elements, such as a block of assessment items or in-game tasks, $\hat{\xi}_e$ is the estimated relative intercept of Element e in a block. $\hat{\alpha}_v$ is the intercept of the vth Variable. Using content of Table 5 as an example, for Element $I1$, which is a test item administered at both pretest and posttest, Element $I1$ gets the same predicted random effect $\hat{\xi}_e$, irrespective of whether the testing occasion is pretest or posttest. $\hat{\alpha}_{Pre}$ is the estimated intercept for Variable *Pre* (the average intercept for all pretest items), $\hat{\alpha}_{Pst}$ is the estimated intercept for Variable *Pst* (the average intercept for all posttest items), and $\hat{\alpha}_{K1}$ is the estimated intercept for a game-based indicator named $K1$. Each intercept can be understood as inversely related to a corresponding difficulty. We refer interested readers to Reckase (2009) for detailed discussions on multidimensional intercept and difficulty terms.

Adding explanatory predictors. We can regress the random terms, η_p and ξ_e, on explanatory predictors, such as design variables and covariates, to answer substantive research questions. Considering the case of a single predictor x_p for Person p and another predictor z_e for Element e, we have the following latent regression equations:

$$\eta_p = \beta_0 + \beta_1 x_p + \epsilon_p, \tag{3}$$

$$\xi_e = \gamma_0 + \gamma_1 z_e + \epsilon_e, \tag{4}$$

where the βs and γs are the latent regression coefficients, and ϵ_p and ϵ_e are the random effects. Substituting Equations 3 and 4 into Equation 2, we obtain

$$z_{vep} = \alpha_v + (\gamma_0 + \gamma_1 z_e + \epsilon_e) + (\beta_0 + \beta_1 x_p + \epsilon_p) \tag{5}$$

or more compactly,

$$z_{vep} = \alpha_v + (\boldsymbol{\gamma}' \boldsymbol{z}_e + \epsilon_e) + (\boldsymbol{\beta}' \boldsymbol{x}_p + \epsilon_p). \tag{6}$$

Adding site dummy variables. Note that Equation 2 does not reflect the nesting of individuals within sites. Suppose there are S sites, one way to account for

such nesting is to regress the person-specific random term on $S-1$ school dummy variables (e.g., on $d_2, \ldots, d_S$ assuming School 1 is the reference school). Now, the person latent variable for Person p in Site s is written as

$$\eta_{ps} = \mu_s d_s + \epsilon_{ps} = \mu_s + \epsilon_{ps}, \tag{7}$$

where ϵ_{ps} becomes Person p's deviation from Site s' intercept μ_s.

The Full Model

We have now discussed key components of a cross-classified IRT model that can jointly model game-based indicator data and assessment data. Following Huang and Cai (2024)'s descriptions, the full model has two main parts: (a) the latent structural model, which examines the relationships between different predictors and the latent variables, and (b) the measurement model, which examines the relationships between the latent variables and the observed responses.

The latent structural model.

Cluster Side. On the cluster side, we formulate the latent regression equations as follows.

$$\boldsymbol{\eta}_{ps} = \boldsymbol{\theta}_{ps} + \mu_s, \tag{8}$$

$$\boldsymbol{\theta}_{ps} = \boldsymbol{B}\boldsymbol{x}_{ps} + \boldsymbol{\epsilon}_{ps}. \tag{9}$$

In Equation 8, $\boldsymbol{\eta}_{ps}$ decomposes into a vector of person-specific latent variables $\boldsymbol{\theta}_{ps}$ for Person p in Site s and an intercept μ_s associated with Site s. Equation 9 shows that we can regress the latent variables on external predictors. Specifically, $\boldsymbol{x}_{ps}$ contains predictors specific to Person p in Site s, and $\boldsymbol{B}$ contains the regression coefficients associated with the person-level predictors. $\boldsymbol{\epsilon}_{ps}$ contains the random terms, and $\boldsymbol{\epsilon}_{ps} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_{Person})$.

If all game-based indicators are designed to measure a general game-based performance, and if all assessment items are administered at two time points (e.g., pretest and posttest), $\boldsymbol{\theta}_{ps}$ becomes a three-dimensional vector. The first element of this vector represents Person p's game-based performance. The second element represents Person p's knowledge measured at the pretest. The third element represents Person p's knowledge measured at the posttest. Equation 8 can be expanded as follows:

$$\boldsymbol{\eta}_{ps} = \begin{bmatrix} \mu_s + \theta_{ps,gameplay} \\ \mu_s + \theta_{ps,pretest} \\ \mu_s + \theta_{ps,posttest} \end{bmatrix}. \tag{10}$$

With pretest and posttest administrations, we might also want to measure the extent of change that has occurred between the two assessments as an direct indicator of learning. To do so, we can introduce a latent change parameterization (see Cai & Houts, 2021; McArdle, 2009).

The core concept underlying a latent change parameterization is as follows. With two time points, we include an additional latent variable, denoted as $\theta_{p\Delta}$, at Time 2. This latent variable represents a difference or change, which is added to the overall

status measured at Time 1. A more general version is discussed in Cai and Houts (2021).

If we let η_{p1} denote Person p's measured knowledge at Time 1 (pretest) and η_{p2} denote this person's measured knowledge at Time 2 (posttest), we can re-express η_{p2} as $\eta_{p2} = \eta_{p1} + \theta_{p\Delta}$. When we rearrange this equation, we can see that $\theta_{p\Delta} = \eta_{p2} - \eta_{p1}$ indeed represents a difference in Person p's knowledge measured at two time points. Based on variance algebra, the variance of the latent change variable $\theta_{p\Delta}$ is typically smaller than the variance of a nonchange latent variable like η_{p1}.

To connect our discussion on latent changes to Equation 8, we reexpress the third element of $\boldsymbol{\theta}_{ps}$ to include a time specific effect $\theta_{p\Delta}$. As discussed, $\theta_{p\Delta}$ represents a latent change from the baseline measured at the pretest, with estimable mean and variance. The expanded Equation 8 becomes

$$\boldsymbol{\eta}_{ps} = \begin{bmatrix} \mu_s + \theta_{ps,gameplay} \\ \mu_s + \theta_{ps,baseline} \\ \mu_s + \theta_{ps,baseline} + \theta_{p\Delta} \end{bmatrix}. \tag{11}$$

In Equation 11, the time-specific effect $\theta_{p\Delta}$ is a latent change that captures the part of individual knowledge or performance measured at Time 2 that is not identical to the same individual's knowledge or performance measured at Time 1. When Time 1 corresponds to the pretest administration and Time 2 to the posttest administration, $\theta_{ps,baseline}$ reflects the baseline status or performance measured at the pretest, and $\theta_{p\Delta}$ reflects the change from pretest to posttest.

Block Side. The blocks are uncorrelated with the clusters. The latent regression equation associating Element e's predictors and block-side latent variables is

$$\boldsymbol{\xi}_e = \boldsymbol{\Gamma} \boldsymbol{z}_e + \boldsymbol{\epsilon}_e. \tag{12}$$

In Equation 12, $\boldsymbol{\Gamma}$ contains the regression coefficients. $\boldsymbol{z}_e$ is a vector containing the predictors. $\boldsymbol{\epsilon}_e$ is a vector of random effects following $MVN(\mathbf{0}, \boldsymbol{\Sigma}_{Block})$. With the data structure demonstrated in Table 5, each element is essentially associated with only one block—either the gameplay or the assessment block.

The measurement model. We assume a 1PL measurement model, with binary observed responses denoted by y_{veps}. One could explore the potential uses of more complex models if the data support such exploration. With a 1PL model, $y_{veps} = 1$ indicates a correct response or the presence of a game-based pattern, whereas $y_{veps} = 0$ indicates an incorrect response or the absence of a game-based pattern. The conditional response probability is

$$P(y_{veps} = 1 \mid \boldsymbol{\eta}_{ps}, \boldsymbol{\xi}_e) = \frac{1}{1 + \exp\left[-(\alpha_v + \boldsymbol{\lambda}_1' \boldsymbol{\eta}_{ps} + \boldsymbol{\lambda}_2' \boldsymbol{\xi}_e)\right]}. \tag{13}$$

In Equation 13, α_v is Variable v's intercept. $\boldsymbol{\lambda}_1$ is a vector that contains Variable v's slopes on $\boldsymbol{\eta}_{ps}$, where $\boldsymbol{\eta}_{ps}$ is a vector of latent variables for Person p in Site s. $\boldsymbol{\lambda}_2$ is a vector that contains Variable v's slopes on $\boldsymbol{\xi}_e$, where $\boldsymbol{\xi}_e$ is a vector of latent variables for Element e. With a 1PL model, the slopes across observed variables loading on the same latent variable are constrained to be equal.

Empirical Data Analysis

Now that we have discussed the data structure and the general modeling framework, we present our proposed application using cross-classified IRT modeling and highlight its advantages over three alternative approaches. We apply each of the approaches to a data set collected from a large-scale RCT of math games.

We include the alternative approaches for illustrative purposes and for responding to a potential inquiry from general researchers. If simpler methods like correlational analysis, multiple linear regression, or more advanced ones like SEM using aggregated outcomes, produce similar patterns of findings regarding the intervention effect and the relationship between gameplay and assessments, why use cross-classified IRT modeling?

Data Source

The data set used by this paper consists of gameplay and pre-post item response data collected from 1,711 students in 24 schools participating in a multisite RCT of math learning games (Chung et al., 2014). Students were randomly assigned to the treatment condition ($n_{treatment} = 873$) and the control condition ($n_{control} = 838$) at the classroom level within each school. The treatment group played four games about rational numbers. The control group played four games about solving equations. Both groups received the same amount of instructional time. All eight games were developed with the same development process and personnel (i.e., the same team developing the knowledge specifications and the same team developing the games). The RCT met the group design standards of What Works Clearinghouse (2015). Its study protocol was reviewed and approved by the University of California, Los Angeles (UCLA) Institutional Review Board (IRB).

Game of Interest

The analysis focuses on *Save Patch*, an educational game designed to target concepts of fraction addition. Figure 2 shows annotated screenshots of two tasks in the game. The top screenshot shows an earlier task that targets the addition of whole units. The bottom screenshot shows a later task that targets the addition of fractions.

As shown in Figure 2, in each task, students are presented with a grid and rope pieces that have the length of one unit or fractions of a unit (e.g., halves). Students attempt each task by dragging various rope pieces to the signposts on the grid to form a path. The path is then followed by the game character. The objective is to use ropes with the correct lengths and guide the character from the starting location to the target location. The addition of the rope pieces follows the rules of fraction addition. For example, only rope pieces with the same unit or part of an unit (same denominator) can be added.

Assessment of Fractions Knowledge

The assessment data contain students' responses to 10 dichotomously scored items used for the pre- and post-assessments. The items target the meanings of the

Figure 2. Two tasks in *Save Patch*. [Color figure can be viewed at wileyonlinelibrary.com]

unit, denominator, numerator, and addition of fractions. Figure 3 shows three sample items. Past research has shown that the assessment has high technical qualities based on traditional metrics, including classical item statistics. For more information on its development and validation, please refer to Vendlinski et al. (2010) and Chung et al. (2014).

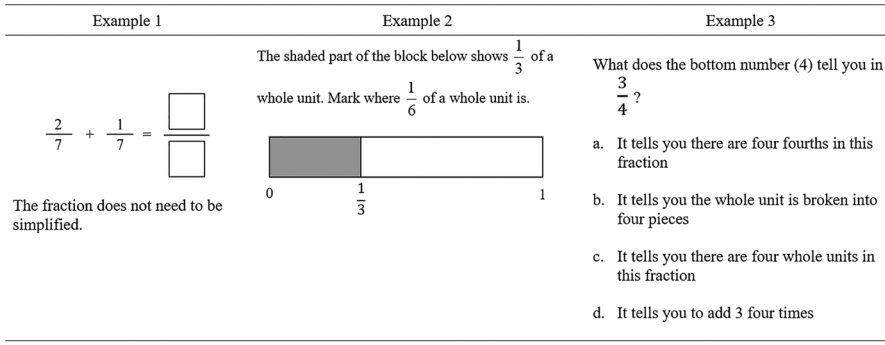

Figure 3. Example assessment items.

Game-Based Indicators of Misconceptions about Fractions

The gameplay indicator data contain students' responses to nine binary indicators of misconceptions, derived using process data from 27 in-game tasks. The choice of these indicators was driven by the set of knowledge specifications central to learning rational numbers (Vendlinski et al., 2010). The same set of knowledge specifications also underlay the development of the game and the pre- and post-assessment items. The development of these indicators was influenced by specifications of the conceptual assessment framework within the evidence-centered design framework (Kerr and Chung, 2012a; Mislevy et al., 2012). The development also heavily relied on (a) the moment-to-moment information in the gameplay process data, (b) how well the in-game interactions addressed cognitive demands related to fractions knowledge, and (c) the degree to which the gameplay data captured cognitively meaningful interactions (Kerr, 2014, p. 90). For more information on development and validation, please refer to Kerr (2014) and Kerr and Chung (2012a, 2012b).

Table 6 presents the nine indicators and their definitions. For this paper, only data based on students' first submissions are used. The first submission window starts when each task begins and ends when a student clicks on the submit button and observes whether or not the submitted solution leads the game character to the target location.

Core Research Questions

We address two research questions that are often asked in RCTs of learning games. The first question concerns evaluating the effectiveness of the intervention in promoting understanding of fraction concepts. The second question concerns how students' performance in the game relates to their performance on traditional assessments.

RQ1. **Treatment Effect**: To what extent do students in the treatment group, who are assigned to play *Save Patch*, differ from students in the control group in their knowledge of fractions as measured by the pre- and post-assessments?

Table 6
List of Game-Based Indicators of Misconceptions about Fractions

No.	Misconception	Indicator	Definition
1	Avoiding math	Placed everything in order	Used all resources in the order that they were presented in a task.
2	Unitizing error	Saw as one unit	Saw the entire grid as one unit.
3		Saw as wholes	Saw each fractional piece as a whole unit.
4	Partitioning error	Counted hash marks	Appeared to count the hash marks on the grid to determine the denominator.
5		Counted hash marks and posts	Appeared to count the hash marks and posts on the grid to determine the denominator.
6	Unitizing and partitioning error	Saw as one unit and counted hash marks	Saw the entire grid as one unit but also appeared to count the hash marks on the grid to determine the denominator.
7		Saw as one unit and counted hash marks and posts	Saw the entire grid as one unit but also appeared to count the hash marks and posts on the grid to determine the denominator.
8	Iterating error	Wrong numerator	Added the wrong number of fractional units.
9	Converting to wholes error	Saw as mixed number	Saw the solution as a mixed number and tried to add a whole unit and a fractional unit without converting everything to have the same denominator.

RQ2. **Relationship between Gameplay and Assessments**: To what extent is students' game-based performance (e.g., misconception) related to their knowledge of fractions?

Data Analyses

We compare our proposed approach, cross-classified IRT modeling, to three alternative approaches used for summarizing and analyzing gameplay and assessment data. By juxtaposing these approaches, we discuss the extent to which each approach addresses the two research questions, discuss key model estimates, and argue for the application of cross-classified IRT modeling.

The four approaches proceed from simpler techniques to more advanced ones, including (a) basic descriptive statistics and correlations, (b) multiple linear regression, (c) structural equation modeling (SEM), and our proposed approach (d) cross-classified IRT modeling. For (a)-(c), we used assessment sum scores and aggregated game-based indicators. For (b)-(d), we created school dummy variables (school-level fixed effects) to account for the nesting of students within schools and the potential impact of school-level differences on individuals' outcomes.

Why We Used School Fixed Effects Rather than Random Effects (Variability)?To account for the nesting of individuals within schools and school-level differences, there are two common approaches: (a) random effects or multilevel models and (b) fixed effects models that include a dummy variable for each school. The choice between fixed effects and random effects depends on the underlying assumptions about the nature of these effects and aims of the study. With the random effects approach, we can use multilevel modeling to allow for varying school-specific intercepts (random intercepts) and further examine varying treatment effects across schools (random coefficients).

In our case, however, we are not interested in modeling variability, whether it is school-level variability or variability (heterogeneity) in treatment effects across schools, an analysis that would be useful in substantive research. Instead, our objective is to enable a gradual transition from building a simpler model, such as a multiple linear regression model, to building more complex (latent variable) models covered in later sections. Modeling school fixed effects achieves this objective while providing one way to control for nesting and school-level differences. Therefore, we include school-specific dummy variables in examples using multiple linear regression, SEM, or cross-classified IRT modeling.

Summary of Results. Table 7 summarizes results from all four analyses. Please note that the effect sizes, along with other results, are provided to show a consistent overall trend (e.g., positive intervention effect). Because values displayed in Table 7 were obtained using different analytic approaches, and each approach used different scaling and parameters, values generated by different analyses, including the effect sizes, are not comparable.

We also note that our proposed approach using cross-classified IRT modeling was able to directly examine the effect of the game-based intervention on the pretest-to-posttest change while making use of item-level assessment data and nonaggregated game-based indicator data.

Table 7
Summary of Results

RQ1. Treatment Effect		
Approach	Effect size	Effect size is relative to
Descriptive statistics	.18	Pooled variance of observed posttest sum scores
Multiple regression	.19	Unit variance of observed posttest sum scores
SEM[a]	.29	Unit variance of posttest latent variable
Cross-classified IRT	.26	Estimated variance of pre-to-post latent change variable

RQ2. Relationship between Gameplay and Assessments		
With observed variables only		
Correlational analysis	1.	Proportion of in-game tasks missed negatively correlated with pretest sum scores ($\rho = -.27$, $p < .001$) and posttest sum scores ($\rho = -.33$, $p < .001$).
	2.	Average number of misconceptions negatively correlated with pretest ($\rho = -.46$, $p < .001$) and posttest sum scores ($\rho = -.47$, $p < .001$).
Multiple linear regression	1.	N/A. The treatment indicator and game-based indicators could not be included in the same model. A separate model with data from only the treatment group is needed to gauge the relationship.
With observed and latent variables		
SEM	1.	The gameplay latent variable, measured through average misconception and proportion of in-game tasks missed, had a negative impact on the posttest ($\hat{\beta} = -3.33$, $SE = .51$, $p < .001$).
	2.	The gameplay and pretest latent variables were negatively correlated ($\hat{\phi} = -.20$, $SE = .03$, $p < .001$).
	3.	With the SEM framework, alternative analyse (e.g., mediation) may also be performed.
Cross-classified IRT	1.	The game-based misconception latent variable was negatively correlated with the baseline latent variable ($\hat{\sigma} = -.64$, $SE = .07$).
	2.	The game-based misconception latent variable was negatively correlated with the pre- to post-latent change variable ($\hat{\sigma} = -.15$).

Note. Each treatment effect is a form of standardized estimate. Each approach uses different scaling and parameters. Values listed in this table are meant to show a general pattern or trend in results (e.g., positive treatment effects). These values are not comparable.
SEM: structural equation modeling.

Table 8
Pretest and Posttest Assessment Sum Scores by Experimental Conditions

Condition	Time	*n*	*M*	*SD*
Control	Pretest	801	4.57	2.64
	Posttest	774	4.80	2.79
Treatment	Pretest	851	4.52	2.68
	Posttest	801	5.33	2.95
Cohen's *d* for posttest		.18		

Note. The total sample size is 1,711 across 24 schools, with 873 students assigned to the treatment condition and 838 assigned to the control condition.

Descriptive statistics and correlational analysis. Descriptive statistics is a typical starting point for understanding both assessment and gameplay data. These statistics include numerical summaries such as the mean and standard deviation for each variable.

RQ1. Treatment Effect. Table 8 presents the sample size excluding missing data (*n*), mean (*M*), and standard deviation (*SD*) of the pretest and posttest sum scores for the treatment and control groups.

We make two observations from results presented in this table. First, the treatment and control groups had similar mean pretest scores and the standard deviations of the scores ($M_{control} = 4.57$, $SD_{control} = 2.64$; $M_{treatment} = 4.52$, $SD_{treatment} = 2.68$). The similarity in pretest scores suggests that the randomization process succeeded in creating two groups with similar levels of prior knowledge as measured by the pretest. Second, following the game-based intervention, the mean posttest score of the treatment group was .53 points higher than that of the control group, indicating a difference of .18 pooled posttest standard deviation. The second observation suggests that the learning game helped enhance students' knowledge of fractions. However, a difference derived from the posttest data alone offers limited insights, as it does not account for students' pretest performance, their performance in the game-based intervention, and the complexities introduced by the multisite design of the RCT. We can address these limitations by using a more sophisticated approach, as discussed in later sections.

Here, we also include two aggregated game-based indicators: the proportion of unique in-game tasks missed and the average frequency of misconceptions. These indicators are used in subsequent analyses, except for cross-classified IRT modeling where nonaggregated indicators can be used. Table 9 shows the descriptive statistics for the two aggregated indicators.

The rationale for choosing the two indicators is twofold. First, the proportion of in-game tasks missed (or conversely, tasks played) represents count- or proportion-based indicators commonly used in existing literature to gauge in-game performance or progress. The preference for this type of indicators may be attributed to their straightforward definition and ease of computation. In our case, the more unique tasks players attempted, the further they progressed in the game, and likely the more

Table 9
Descriptive Statistics of Two Aggregated Game-Based Indicators

Variable	*M*	*SD*	Min.	Max.
Average frequency of misconceptions	.29	.16	.00	.71
Proportion of unique tasks missed	.06	.16	.00	.93

Note. Results are based on data from 826 students with analyzable gameplay data. The average frequency of misconceptions is computed based on values summed across nine indicators and averaged across in-game tasks played by each individual.

Table 10
Correlations between Game-Based and Assessment Measures (n = 786-826)

	1	2	3
1. Prop. unique tasks missed	-		
2. Avg. frequency of misconceptions	.29***	-	
3. Pretest sum scores	−.27***	−.46***	-
4. Posttest sum scores	−.33***	−.47***	.74***

Note. ***$p < .001$.

they knew or learned. To compute the proportion, we counted the number of in-game tasks not played by an individual and divided this count by 27, which was the total number of unique tasks. Of note, the tasks were organized into stages, and each stage had its gameplay flow and fraction knowledge specifications[1] (Center for Advanced Technology in Schools, 2012).

Second, the average frequency of misconceptions was calculated based on the nine specific misconception indicators shown in Table 6. When computing the average frequency of misconceptions for each student, we noted that each of the nine binary indicators denoted the presence or absence of a particular fraction-related misconception. We aggregated these indicator values up to the individual level by summing the number of 1s across the nine indicators and then averaging this sum by the number of tasks played by an individual. The aggregated indicator provides an overall measure of the frequency of misconceptions related to adding fractions (in students' initial submissions).

RQ2. Relationship between Gameplay and Assessments. We can display gameplay and assessment data graphically in more than one dimension to inspect potential relationships among the variables. On the other hand, creating graphs becomes more complex and cumbersome as the number of variables grows. In such cases, correlational analysis provides one way to summarize the pairwise relationships, including their directions and strengths, observed through visualization.

Table 10 shows the pairwise nonparametric correlations (Spearman's rho) between the aggregated game-based indicators and assessment scores. The magnitudes of the correlations between indicators and assessment scores were weak to moderate. Specifically, students who missed more in-game tasks tended to exhibit more

misconceptions in their first submissions ($\rho = .29$, $p < .001$) while having lower pretest ($\rho = -.27$, $p < .001$) and posttest sum scores ($\rho = -.33$, $p < .001$). Students exhibiting more misconceptions tended to have lower pretest ($\rho = -.46$, $p < .001$) and posttest sum scores ($\rho = -.47$, $p < .001$).

The findings presented above are consistent with our expected relationships between misconceptions detected during gameplay and assessment scores. By inspecting descriptive statistics and pairwise correlations, we have two findings relevant for answering the core research questions. First, we found a difference of.18 pooled standard deviations based on the posttest sum scores, which falls within the typical range observed in studies focusing on game-based (math) learning.[2] While this magnitude may appear small when compared to the $> .40$ hinge point (Hattie, 2023; Mayer, 2019), we caution against overreliance on using a fixed threshold to evaluate instructional effectiveness universally, as argued by Hattie (2023, p. 30) and Kraft (2020). Second, we observed negative relationships between game-based nonpositive behaviors (tasks missed and misconceptions) and higher assessment scores. These two findings offer preliminary insights into whether the game-based intervention positively impacted students' knowledge of fractions and how students' in-game performance related to their assessment performance.

However, insights gleaned from descriptive statistics and pairwise correlations are limited and fragmented. First, the assertion about the treatment effect relied on comparing the experimental groups' posttest sum scores without considering the nesting of individuals within schools. Second, correlations were only about pairwise associations. These concerns can be addressed by using a more advanced method that accommodates multiple variables or predictors, and models either school fixed effects or random effects to account for nesting. Lastly, findings were derived from separate analyses and were not integrated, as they would be in a unified modeling framework. For example, findings from the correlational analysis were not considered in the calculation of the noted.18 difference. In the next section, we use a multiple linear regression model with added school dummy variables to address some of the concerns discussed thus far.

Multiple linear regression. With a multiple linear regression model, we can examine the relationships between students' posttest scores and other variables, including pretest sum scores and the treatment assignment indicator.

Model Specification. We specified the multiple regression model with school fixed effects as follows. We regressed the dependent variable, posttest sum scores, on pretest sum scores, treatment condition indicator, and 23 school dummy variables with School 1 as the reference school [$F(25, 1504) = 90.74$, $p < .001$, $R^2 = .60$]. The inclusion of school dummy variables allowed each school to have its own intercept. Each coefficient associated with a dummy variable denoted a shift in the intercept for a specific school in comparison to the reference school.

RQ1. Treatment Effect. Results presented in Table 11 showed a positive estimated coefficient on the treatment assignment indicator ($\hat{\beta} = .19$, $SE = .04$, $t = 4.74$, $p < .001$) after controlling for pretest sum scores and school-level differences. In simpler terms, we observed a positive estimated effect of the game-based

Table 11
Multiple Regression Results Using Posttest Sum Scores as the Dependent Variable

	Estimate	*SE*	*t*
Intercept	.23*	.09	2.46
Pretest sum scores	.73***	.02	41.91
Treatment	.19***	.04	4.74

Note. Pretest and posttest sum scores are mean-centered and scaled by one *SD*. The estimated coefficients on the school dummy variables are not shown in the table; these estimates range between −.68 ($SE = .12$) and −.07 ($SE = .11$). *$p < .05$; ***$p < .001$.

intervention, which was about .19 *SD* increase in posttest sum scores for the treatment group compared to the control group. Note that the effect of .19 was obtained after standardizing the pretest and posttest sum scores. Each sum score variable was standardized by subtracting from it its sample mean and dividing it by its standard deviation.

Given a 10-item assessment that was dichotomously scored, it may also be of interest to examine the unstandardized coefficient on the treatment assignment indicator. With the same set of predictors, the estimated coefficient became $\hat{B} = .55$ ($SE = .12$, $t = 4.74$, $p < .001$). In other words, on average, students who played the game about adding fractions ($Treatment = 1$) scored .55 points higher in their posttest sum scores or answered .55 more items correctly in their posttest, compared to students who did not play the game about adding fractions ($Treatment = 0$).

RQ2. Relationship between Gameplay and Assessments. We did not include any game-based indicator into the model, and as a result, we did not have results pertaining to the second research question.

Why didn't we include any game-based indicators? Recall that roughly half of the sample was assigned to the treatment condition and asked to play the game *Save Patch* about adding fractions, and there were no *Save Patch* data for the other half of the sample. If we were to include any game-based indicator into the multiple regression model, we would lose half of our data—data of students in the control condition, making it impossible to estimate the coefficient on the treatment assignment indicator. Said differently, it would be impossible to estimate the treatment effect given the exclusion of the control group's data. An alternative is to create two separate models: one considering the treatment assignment indicator and another considering the game-based indicators.

Multiple linear regression has several advantages over descriptive statistics and pairwise correlations. One advantage is its ability to control for multiple predictors, such as students' pretest sum scores, and obtain a more precise estimate of the treatment effect. Another advantage is its ability to account for school-level differences by adding school dummy variables (school-level fixed effects).

However, with the aforementioned multiple linear regression model, we cannot simultaneously examine the effectiveness of the game-based intervention and the relationship between gameplay and assessments. The model also limits our ability

to pose additional questions or conduct further analyses. For example, we cannot investigate how student-specific factors affect assessment performance by regressing a variable representing the overall assessment performance on background variables. Meanwhile, we cannot explore how elements of game design affect in-game performance by regressing a variable representing the overall in-game performance on indicators of game-specific cognitive or mechanic features. To overcome these limitations, we turn to latent variable models with added school dummy variables. We begin with the use of structural equation modeling.

Structural equation modeling (SEM). Researchers may use structural equation models to jointly analyze assessment sum scores and aggregated game-based indicators, moving from the modeling of observed variables to that of both observed and latent variables. We include this SEM example for illustrative purposes only, showing a latent variable modeling approach different from cross-classified IRT modeling.

Before we specify the model, we note two potential modeling limitations associated with using SEM to jointly analyze aggregated gameplay and assessment outcomes. First, joint modeling presented in this SEM example relies on treating the two game-based indicators as outcome variables instead of covariates, and on assuming all outcome variables to be conditionally normal. Second, in this SEM example, the pretest and posttest latent variables are single-indicator latent variables. This setup assumes that the a single indicator (e.g., pretest sum score) fully reflects the measured phenomenon (e.g., pretest performance) without measurement error (Raykov & Marcoulides, 2006). Both the assumption of normality and the assumption of no measurement error likely do not hold in this specific context. Therefore, we caution readers who still want to apply SEM to the analysis of aggregated gameplay and assessment data to stay vigilant about the single-indicator situation and to conduct thorough checks for possible violations of model assumptions.

Model Specification. The specified model had three latent variables. The first latent variable, labeled PRETEST, represented pretest performance. The second, labeled POSTTEST, represented posttest performance. The third latent variable, labeled GAMEPLAY, reflected nonpositive gameplay performance. PRETEST and POSTTEST were single-indicator latent variables. For example, the PRETEST latent variable predicted only one observed outcome variable, the pretest sum score. Similarly, the POSTTEST latent variable predicted only one observed outcome variable, the posttest sum score. When connecting an observed outcome variable with a single-indicator latent variable, we fixed the observed variable's factor loading to 1 and its error variance to 0. We also fixed the constant intercept term of the observed variable to be the observed variable's mean.

The third latent variable, GAMEPLAY, predicted two observed outcome variables: the proportion of in-game tasks missed and the average frequency of misconceptions. GAMEPLAY reflected nonpositive performance because the two game-based indicators were conceptually and empirically shown, for example, through pairwise correlations, to be negatively related to higher test scores. In the model, we fixed the factor loading of the misconception indicator to 1, leaving the other indicator's

factor loading freely estimated. We also freely estimated the error variances associated with both game-based indicators.

Among the latent variables, we regressed POSTTEST on PRETEST and GAMEPLAY, respectively. We estimated the variance associated with each latent variable or its disturbance, and we also estimated the covariance between PRETEST and GAMEPLAY. We assumed that the means of the latent variables were 0.

To estimate the treatment effect, we regressed POSTTEST on the treatment assignment indicator. We estimated both unstandardized and standardized treatment effects. The standardized solution was obtained by invoking the *SS* command in LISREL 12 (Jöreskog & Sörbom, 2023), which standardized all latent variables.

To account for school-level differences, we also regressed POSTTEST on 23 school dummy variables, using School 1 as the reference school. As explained at the start of the "Data Analyses" section, we are not interested in exploring the variability of treatment effects across schools. Instead, we include school dummy variables as a method to address the nesting of students in schools. Our focus remains on the joint analysis of gameplay and assessment data.

Note that the specified SEM model is one of the possible models for examining the relationship between gameplay and assessments. An alternative model could involve GAMEPLAY acting as a mediator, in which case the interpretation of the treatment effect, as indicated by the coefficient on the treatment assignment indicator, can get more complicated. The model discussed in this section is used to build on the multiple regression analysis, where we regressed posttest sum scores on pretest sum scores, the treatment assignment indicator, and school dummy variables.

With the above specifications, we estimated the model with full information maximum likelihood in LISREL 12 (Jöreskog & Sörbom, 2023).

RQ1. Treatment Effect. We examined the coefficient on the treatment assignment indicator to gauge the treatment effect. The standardized estimated coefficient was .29 ($SE = .01$, $t = 24.57$, $p < .001$), suggesting a positive impact of the game-based intervention on the latent posttest performance, after considering the relationships between POSTTEST and other variables of interest (e.g., PRETEST).

Compared with the effect obtained in the multiple regression analysis, this effect of .29 was estimated while considering the following information: (a) students' pretest performance; (b) information or process evidence from the intervention (game) itself; and (c) the interplay between game- and assessment-based outcomes. With this joint analysis, our claim about the effectiveness of the game-based intervention on student knowledge outcomes becomes grounded in multiple sources of information. Such a joint analysis differs from the previous analyses that solely relied on assessment data.

RQ2. Relationship between Gameplay and Assessments. There were two findings on the relationship between gameplay and assessments. First, GAMEPLAY, a latent variable measured through the average frequency of misconception and the proportion of in-game tasks missed, had a negative impact on POSTTEST ($\hat{\beta} = -3.33$, $SE = .51$, $t = 6.50$, $p < .001$). This result suggests that students with better

performance in the game, such as exhibiting fewer misconceptions, tended to score higher on the posttest, after accounting for influences of other variables of interest on POSTTEST. Second, GAMEPLAY and PRETEST were negatively correlated ($\hat{\phi} = -.20$, $SE = .03$, $t = 7.13$, $p < .001$), suggesting that students who scored higher on the pretest or started with greater prior knowledge of fractions tended to perform better in the game, and hence having fewer misconceptions and fewer missed tasks.

The patterns we observed with the SEM approach, including the sign and magnitude of the treatment effect and how gameplay related to assessments, were consistent with patterns seen in previous analyses. First, we observed positive treatment effects across all analyses discussed thus far. Second, in both correlational and SEM analyses, we observed an overall negative association between nonpositive gameplay, such as showing misconceptions, and better assessment performance.

We also note the methodological differences in what model parameters were involved and how the effects were estimated. For instance, the SEM-based effect size was obtained after standardizing the latent variables and thus was relative to the unit variance of a latent variable, not an observed variable. In comparison, the effect size computed using descriptive statistics was based on observed posttest sum scores exclusively. The purpose of discussing these effect sizes is not to imply their comparability. Rather, we discuss them to discern the overall trend in how the game-based intervention influenced student knowledge outcomes.

By using a structural equation model, we can address a key drawback of the multiple linear regression approach: the inability to jointly include gameplay and assessment data. With multiple regression, we cannot include gameplay data with the treatment assignment indicator in a single regression model without losing roughly half of the sample. Consequently, the multiple regression model that includes the treatment assignment indicator lacks any variables related to gameplay. What sets a more advanced approach like SEM apart from earlier, simpler methods is its ability to simultaneously analyze gameplay and assessment data while also having the flexibility to handle certain data complexities and enable researchers to investigate secondary hypotheses.

In addition to the two modeling limitations mentioned earlier, another limitation of the SEM approach lies in the use of aggregated data. Item-level assessment data were aggregated into sum scores, and task-specific game-based indicators were aggregated into individual-specific values. This aggregation procedure, particularly concerning gameplay data, overlooks the varying characteristics of the in-game tasks. In *Save Patch*, the tasks are designed to gradually introduce key concepts and skills about adding fractions as players progress through the game (Center for Advanced Technology in Schools, 2012). For example, players first tackle the addition of whole units before delving into the addition of fractions with the same denominator but different numerators, and ultimately, fractions with varying denominators and numerators. The varying characteristics of the in-game tasks can affect how players engage with and perform in the game, encouraging us to explore an alternative that leverages nonaggregated assessment and gameplay data.

Cross-classified item response theory modeling. We have analyzed assessment and gameplay data using three approaches: descriptive statistics and correlational

analysis, multiple linear regression, and SEM. For each approach, we discussed the findings and limitations, highlighting the additional insights gained from using a more flexible and complex approach.

With descriptive statistics and correlational analysis, we obtained preliminary insights into the extent to which the game-based intervention improved student knowledge outcomes (e.g., posttest sum scores). However, this approach had several limitations, including (a) the inability to adjust for students' pretest or baseline performance when comparing posttest performance between the treatment and control groups; (b) the inability to account for data complexities, such as the nesting of students within schools; and (c) the absence of integrated findings due to the lack of a unified and flexible modeling framework that jointly analyzes data from different sources, such as gameplay and assessments.

With multiple linear regression, we addressed two limitations associated with the use of descriptive statistics and correlational analysis. First, we compared students' posttest sum scores while also considering their pretest sum scores. Second, we included school dummy variables into the regression model (school fixed effects) as a means to account for the nesting of students within schools and the impact of school-level differences on student outcomes. However, one serious drawback remained: the inability to jointly analyze gameplay and assessment data. Including any gameplay data into this regression model would result in the loss of approximately half of the data, specifically data of students in the control condition who did not play the game, and the loss of the ability to estimate the treatment effect.

With SEM, we included both aggregated assessment sum scores and game-based indicators in the same model while retaining the ability to address complexities such as the nesting of students within schools. While we appreciated the flexibility and the potential for alternative model specifications offered by the SEM framework, we also noted the limitations of a SEM-based approach that relied on aggregated outcomes. Data aggregation, in the context of game-based research, overlooked the varying characteristics of in-game tasks introduced by design.

Furthermore, all three preceding approaches did not directly model changes in students' performance. We would be one step closer to making claims about student learning if we could directly examine changes in performance, such as changes from pre- to post-assessments. The earlier methods, including descriptive statistics, correlational analysis, and multiple linear regression, used posttest sum scores as the outcome, controlling for pretest sum scores when applicable. In the SEM application using aggregated outcomes, we regressed a latent variable representing posttest performance on the treatment assignment indicator. Although SEM, particularly longitudinal SEM using latent-change concepts (e.g., McArdle, 2009), can model latent changes, we did not explore additional structural equation models because we wanted to transition away from using aggregated data, especially aggregated game-based indicator data.

Given the aforementioned considerations, we now present our proposed approach that uses a cross-classified IRT model to jointly analyze nonaggregated gameplay and assessment data.

Structure of Input Data. The general structure of the input data mirrors the structure presented in Table 5 in the "Example of the Combined Data" section. The term "Block" refers to either the assessment block of items or the gameplay block of in-game tasks. Person-specific latent variables are referred to as cluster side latent variables. Item- or game task-specific latent variables are referred to as block side latent variables.

Model Specification. The model is a specific case within the general cross-classified IRT modeling framework described earlier.

The latent structural component of the model had a total of five latent variables. Among these, three were on the cluster side, representing individual game-based misconception, baseline performance, and pretest to posttest change, respectively. We assumed a consistent misconception latent variable because the gameplay duration and exposure for one single game was relatively short. The other two latent variables were on the block side. Specifically, the gameplay block captured varying relative intercepts of in-game tasks, while the assessment block captured varying relative intercepts of pre- and post-assessment items.

We imposed or relaxed constraints on the means, variances, and covariances associated with the latent variables as follows. For the misconception and baseline latent variables, as well as the two block side latent variables, we constrained their means to 0 and variances to 1. We freely estimated (a) the mean and the variance of the latent change variable, (b) the covariance between the misconception and baseline latent variables, and (c) the covariance between the misconception and latent change variables.

To estimate the treatment effect and account for school effects, we regressed the baseline and change latent variables on covariates. Specifically, we regressed both the baseline and change latent variables on the treatment assignment indicator. Also, we regressed the baseline latent variable on 23 school dummy variables, making School 1 the reference school.

We specified the relationships between observed and latent variables as follows, starting with the cluster side latent variables and moving to the block side latent variables. First, the nine game-based indicators loaded on the first latent variable (misconception), and their slopes were constrained to be equal.[3] Second, the two observed variables containing assessment item responses, denoted as pretest and posttest, loaded on the second latent variable (baseline), while the posttest variable also loaded on the third latent variable (latent change). This configuration enabled us to use the third latent variable to capture students' latent change scores. Third, we constrained slopes of the pretest and posttest variables loading on the second latent variable, as well as the slope of the posttest variable loading on the third latent variable, to be equal. Fourth, we constrained the intercepts of the pretest and posttest variables to be equal. Fifth, we constrained slopes of the game-based indicators loading on the fourth latent variable (the gameplay block) to be equal. Lastly, we constrained slopes of the pretest and posttest variables loading on the fifth latent variable (the assessment block) to be equal.

We estimated the specified model using the Metropolis-Hastings Robbins-Monro algorithm implemented in flexMIRT 3.65 (Cai, 2022). The chosen random seed was

Table 12
Estimates and SEs of Group Parameters and Latent Regression Coefficients

Parameter	Game-Based Misconception	Baseline	Latent Change	Block (Tasks)	Block (Items)
Latent mean (*SE*)	.00 (—)	.00 (—)	.11 (.01)	.00 (—)	.00 (—)
Regression coefficient (*SE*)		.08 (.06)	.26 (.03)		
Covariance matrix (*SE*)	1.00 (—)				
	−.64 (.07)	1.00 (—)			
	−.06 (.02)	.00 (—)	.17 (.03)		
	.00 (—)	.00 (—)	.00 (—)	1.00 (—)	
	.00 (—)	.00 (—)	.00 (—)	.00 (—)	1.00 (—)

1010. The number of imputations from the MH step per RM cycle was 1. The dispersion values of the Metropolis proposal densities for the first and second levels of the specified model were 3.0 and 2.0, respectively. The number of Stage I (constant gain) cycles was 10,000, and the number of Stage II (Stochastic EM, constant gain) cycles was 1,000. The chosen scoring method was Expected A Posteriori (EAP). In addition, we enabled options to save the iteration history, estimated model parameters, and individual IRT scale scores. All other estimation settings were the defaults described in Houts and Cai (2020).

Table 12 presents the estimated group parameters and regression coefficients on the treatment assignment indicator. The estimated coefficients on the school dummy variables ranged between $-.44$ ($SE = .14$) and .65 ($SE = 0.11$).

Table 13 presents the estimated parameters associated with the assessment items and the game-based indicators. A game-based indicator can be likened to an item, with its estimated intercept inversely related to a corresponding difficulty. Each indicator's estimated intercept also reflects the degree of prevalence of a specific misconception about fraction addition.

RQ1. Treatment Effect. Based on the estimated regression coefficients presented in Table 12, the control and treatment groups did not differ significantly in their baseline performance measured at pretest ($\hat{\beta} = .08$, $SE = .06$, $p = ns$). The two groups, in comparison, did differ significantly in their latent changes, and there was a positive treatment effect on the latent pretest-to-posttest change ($\hat{\beta} = .26$, $SE = .03$, $p < .05$). Note that this effect size of .26 is relative to the variance of a latent change variable. As discussed in "The Full Model" section, a latent change variable tends to have a variance that is smaller than the variance of a nonchange latent variable.

RQ2. Relationship between Gameplay and Assessments. The covariance matrix presented in Table 12 provides information on the relationship between gameplay and assessments. Game-based misconception was negatively correlated with the baseline performance ($\hat{\sigma}_{base,misc} = -.64$, $SE = .07$). Because the game-based indicators targeted misconceptions about adding fractions, the direction of this correlation aligned with the expectation: students with lower baseline performance tended to exhibit more misconceptions during gameplay. Game-based misconception was also

Table 13
Item Parameter Estimates and SEs

Variable	Label	λ_{misc}	*SE*	λ_{base}	*SE*	λ_{chg}	*SE*	λ_{b1}	*SE*	λ_{b2}	*SE*	Intercept	*SE*
1	EvIO	.47	.02							.87	.05	−3.80	.05
2	SAOU	.47	.02							.87	.05	−3.65	.05
3	SAOUaCHM	.47	.02							.87	.05	−4.71	.08
4	SAOUaCHMaP	.47	.02							.87	.05	−5.95	.15
5	CHMaP	.47	.02							.87	.05	−2.59	.03
6	SwAW	.47	.02							.87	.05	−4.01	.06
7	CnHM	.47	.02							.87	.05	−2.33	.03
8	WrnN	.47	.02							.87	.05	−2.27	.03
9	SAMN	.47	.02							.87	.05	−5.04	.09
10	Pretest Items			1.49	.02			1.36	.03			−0.30	.02
11	Posttest Items			1.49	.02	1.49	.02	1.36	.03			−.30	.02

Note. The nine game-based indicators are abbreviated: EvIO = Everything in order, SAOU = Saw as one unit, SAOUaCHM = Saw as one unit and counted hash marks, SAOUaCHMaP = Saw as one unit and counted hash marks and posts, CHMaP = Counted hash marks and posts, SwAW = Saw as wholes, CnHM = Counted hash marks, WrnN = Wrong numerator, and SAMN = Saw as mixed number. Other abbreviations are *misc* for game-based misconception, *base* for baseline performance, *chg* for latent change, *b*1 for Block 1, the assessment block, and *b*2 for Block 2, the gameplay block.

negatively correlated with the latent change ($\hat{\sigma}_{chg,misc} = -.06/\sqrt{.17} = -.15$). This result suggests that students who exhibited more misconceptions in their gameplay tended to experience a lesser degree of change from pretest to posttest.

These findings have three important implications for learning and instruction. First, information extracted from gameplay process data, such as indicators of misconceptions, can help surface students' prior knowledge and common conceptions concerning the addition of fractions. Second, the presence of misconceptions during gameplay may hinder or limit the extent to which students can make progress in their understanding of adding fractions. Third, these findings can serve as valuable inputs for instructional planning. For example, teachers may adjust subsequent instructions to address common conceptions and misconceptions identified during students' gameplay, with the goal of promoting conceptual change and improving student learning outcomes.

Summary and Discussion

We began this paper by framing the understanding of data from game-based evaluation studies as a sensemaking process. Recognizing the breadth and complexity of this process, we focused on presenting a new application of cross-classified IRT modeling. The application showed how we could use a cross-classified IRT model to jointly analyze item-level data from traditional assessments and nonaggregated gameplay process information summarized by diagnostic indicators, while addressing key aspects of (multisite) evaluation research of educational games. With the latent change parameterization discussed in "The Full Model" section, the proposed application could connect individuals' performance in a digital game with changes in

assessment outcomes, providing one way to directly evaluate the impact of a game-based intervention on learning.

One notable limitation of our paper is the lack of discussions on model fit, especially for parameterized latent variable models. While a major goal of this paper is to demonstrate the utility of a modern and flexible modeling framework, we recognize that it is still important to examine how well a model fits the data before interpreting any estimated parameters, given that inferences drawn from a model are contingent on the quality of the model itself.

Another aspect that we did not discuss is player or learner engagement in games. Engagement in games, or defining, operationalizing, and measuring engagement in educational games, is a complex topic in itself (e.g., Abdul Jabbar & Felicia, 2015; Hookham & Nesbitt, 2019). It remains important to investigate the interplay of engagement in educational games, performance, and learning. We thus invite future research to measure more of the extent to which learners engage in educational games.

Moreover, two preconditions need to be met in order to fully realize the benefits of the modeling application proposed in our paper. The application of cross-classified IRT modeling to integrate game and assessment data places great value on (a) having a set of content and cognitive demand features that guide the development of both learning (e.g., educational games) and assessment systems, and (b) having diagnostic or theory-informed indicators developed using process data. Meeting the first precondition greatly facilitates the second. We believe that these two preconditions, along with having a unified and flexible modeling framework like cross-classified IRT modeling, are crucial for creating a coherent analytical framework that enables us to harness diagnostic information from process data, reduces our reliance on summative assessments, and enhances the instructional process by offering targeted feedback to students and instructors.

Compared to fully data-driven procedures, the shortcoming of the kind of approach advocated in our paper—emphasizing sensemaking and meeting preconditions—is also evident. Sensemaking is a volitional process, where we actively come to understand the differences as well as the interconnectivity embedded in multiple sources of information, amid other complexities. This process is time and labor intensive, requiring us to navigate massive volumes and diverse types of data and distill them into diagnostic, actionable information about effectiveness and areas for improvement. The challenges and requirements posed by this process can be daunting, especially when juxtaposed with high-potential machine learning or artificial intelligence applications, wherein tasks such as indicator development can be relegated.

Nevertheless, we believe that there is value in sensemaking, whether it is for the evaluation of educational games or for the analysis of game-based evaluation data. Developing and executing a well-defined theory of action, as well as meeting preconditions, is an investment that is likely to result in "more efficient analysis and more valid interpretation of the data" (Goldhammer et al., 2021; Lindner & Greiff, 2023; Zumbo et al., 2023, pp. 245-246). Returning to the *frame* argument in sensemaking, we also want to underscore the centrality of the researchers in the scientific discovery process. While data-driven methods are valuable for identifying unseen trends,

researchers play a crucial role in shaping, guiding, and engaging in conversations about the discovery process.

Acknowledgments

Dr. Li Cai's research is partially supported by a grant from the Institute of Education Sciences (R305D210032). The views expressed in this paper belong to the co-authors and do not represent those of the funding agency.

Notes

[1]Of the 826 students with analyzable gameplay data, 704 students (85%) attempted all 27 in-game tasks, and 752 students (91%) attempted at least 20 of the 27 in-game tasks, where the gameplay flow in the first 20 tasks overlapped with the flow in the last 7 tasks. Starting Task 16, the numbers and types of fraction knowledge specifications targeted by each task were also similar. A total of 777 of 826 students (94%) attempted Task 16 or beyond.

[2]Based on a meta-analysis of 24 studies examining the impact of game-based learning on student math achievement, the overall weighted effect size was .13 with an associated 95% confidence interval of [.02, .24] (Tokac et al., 2019). Based on another meta-analysis of 48 studies that examined the impact of educational games on learning outcomes and used games as the only instructional method, the overall weighted effect size was .20 with an associated 95% confidence interval of [.03, .37] (Wouters et al., 2013).

[3]We imposed this constraint considering the small sample size relative to the complexity of the model, especially if all of these slopes were to be freely estimated. If warranted by theoretical and practical considerations, such as having a larger sample size and data conditions that would support the estimation of a more complex model, it is possible to relax this constraint.

References

Abbot, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological Methods & Research*, *29*(1), 3–33.

Abdul Jabbar, A. I., & Felicia, P. (2015). Gameplay engagement and learning in game-based learning: A systematic review. *Review of Educational Research*, *85*(4), 740–779.

Arieli-Attali, M., Ward, S., Thomas, J., Deonovic, B., & von Davier, A. A. (2019). The expanded evidence-centered design (e-ECD) for learning and assessment systems: A framework for incorporating learning goals and processes within assessment design. *Frontiers in Psychology*, *10*.

Baker, E., Chung, G., & Delacruz, G. (2008). Design and validation of technology-based performance assessments. In J. M. Spector, M. D. Merrill, J. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3 edn., pp. 595–604). Lawrence Erlbaum Associates.

Bergner, Y., & von Davier, A. A. (2019). Process data in NAEP: Past, present, and future. *Journal of Educational and Behavioral Statistics*, *44*(6), 706–732.

Blanié, A., Amorim, M.-A., Meffert, A., Perrot, C., Dondelli, L., & Benhamou, D. (2020). Assessing validity evidence for a serious game dedicated to patient clinical deterioration and communication. *Advances in Simulation*, *5*(4), 1–12.

Cagiltay, N. E., Ozcelik, E., & Ozcelik, N. S. (2015). The effect of competition on learning in games. *Computers & Education*, *87*, 35–41.

Cai, L. (2022). flexMIRT®: Flexible multilevel multidimensional item analysis and test scoring. Computer software.

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

Cai, L., & Houts, C. R. (2021). Longitudinal analysis of patient-reported outcomes in clinical trials: Applications of multilevel and multidimensional item response theory. *Psychometrika*, *86*(3), 754–777.

Center for Advanced Technology in Schools (2012). CATS-developed games. (CRESST Resource Paper No. 15). University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). https://cresst.org/publications/cresst-publication-3255/

Chen, F., Cui, Y., & Chu, M.-W. (2020). Utilizing game analytics to inform and validate digital game-based assessment with evidence-centered game design: A case study. *International Journal of Artificial Intelligence in Education*, *30*(3), 481–503.

Chen, Y., Zhang, J., Yang, Y., & Lee, Y.-S. (2022). Latent space model for process data. *Journal of Educational Measurement*, *59*(4), 517–535.

Chia, R. (2000). Discourse analysis organizational analysis. *Organization*, *7*(3), 513–518.

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

Chung, G. K. W. K., & Baker, E. (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).

Chung, G. K. W. K., Choi, K., Baker, E. L., & Cai, L. (2014). The effects of math video games on learning: A randomized evaluation study with innovative impact estimation techniques. CRESST.

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—Designs, methods and solutions*. Springer.

Darling-Hammond, L., Herman, J., Pellegrino, J., Abedi, J., Aber, J. L., Baker, E., Bennett, R., Gordon, E., Haertel, E., Hakuta, K., Ho, A., Linn, R. L., Pearson, P. D., Popham, J., Resnick, L., Schoenfeld, A. H., Shavelson, R., Shepard, A., Shulman, L., & Steele, C. M. (2013). Criteria for high-quality assessment. Stanford Center for Opportunity Policy in Education.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*(4), 533–559.

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, *10*(102), 1–11.

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

Dervin, B. (2003). *Sense-making methodology reader: Selected writings of Brenda Dervin*. Hampton Press.

Ercikan, K., Guo, H., & He, Q. (2020). Use of response process data to inform group comparisons and fairness research. *Educational Assessment*, *25*(3), 179–197.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308.

Foster, N., & Piacentini, M. (Eds.). (2023). *Innovating assessments to measure and support complex skills*. OECD Publishing.

Gane, B. D., Zaidi, S. Z., & Pellegrino, J. W. (2018). Measuring what matters: Using technology to assess multidimensional learning. *European Journal of Education*, *53*(2), 176–187.

Garcia, I., Pacheco, C., Méndez, F., & Calvo-Manzano, J. A. (2020). The effects of game-based learning in the acquisition of "soft skills" on undergraduate software engineering courses: A systematic literature review. *Computer Applications in Engineering Education*, *28*(5), 1327–1354.

Gauthier, A., Corrin, M., & Jenkinson, J. (2015). Exploring the influence of game design on learning and voluntary use in an online vascular anatomy study aid. *Computers & Education*, *87*, 24–34.

Goldhammer, F., Hahnel, C., Kroehne, U., & Zehner, F. (2021). From byproduct to design factor: On validating the interpretation of process indicators based on log data. *Large-scale Assessments in Education*, *9*(1), 20.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*(3), 608–626.

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education*, *91*, 92–105.

Hahnel, C., Jung, A. J., & Goldhammer, F. (2023). Theory matters: An example of deriving process indicators from log data to assess decision-making processes in web search tasks. *European Journal of Psychological Assessment*, *39*(4), 271–279.

Hao, J., Shu, Z., & Davier, A. Von. (2015). Analyzing process data from game/scenario-based tasks: An edit distance approach. *Journal of Educational Data Mining*, *7*(1), 33–50.

Hattie, J. (Ed.). (2023). *Visible learning: A synthesis of over 2,100 meta-analyses relating to achievement*. Routledge.

Hautala, J., Heikkilä, R., Nieminen, L., Rantanen, V., Latvala, J.-M., & Richardson, U. (2020). Identification of reading difficulties by a digital game-based assessment technology. *Journal of Educational Computing Research*, *58*(5), 1003–1028.

Hookham, G., & Nesbitt, K. (2019). A systematic review of the definition and measurement of engagement in serious games. In *Proceedings of the Australasian Computer Science Week Multiconference*, ACSW'19 (pp. 1–10). Association for Computing Machinery.

Houts, C. R., & Cai, L. (2020). *flexMIRT® user's manual version 3.6: Flexible multilevel multidimensional item analysis and test scoring*. Vector Psychometric Group. Software manual.

Huang, S., & Cai, L. (2024). Cross-classified item response theory modeling with an application to student evaluation of teaching. *Journal of Educational and Behavioral Statistics*, *49*(3), 311–341. https://doi.org/10.3102/10769986231193351

Jiao, H., He, Q., & Veldkamp, B. P. (2021). Editorial: Process data in educational and psychological measurement. *Frontiers in Psychology*, *12*, 793399.

Jiao, H., Liao, D., & Zhan, P. (2019). Utilizing process data for cognitive diagnosis. In: M. von Davier & Y. S. Lee (Eds.), *Handbook of diagnostic classification models: Models and model extensions, applications, software packages*. Methodology of Educational Measurement and Assessment (pp. 421–436). Springer.

Jöreskog, K., & Sörbom, D. (2023). LISREL 12. Computer software.

Kerr, D. (2014). Into the black box: Using data mining of in-game actions to draw inferences from educational technology about students' math knowledge.

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*(1), 144–182.

Kerr, D., & Chung, G. K. W. K. (2012b). The mediation effect of in-game performance between prior knowledge and posttest score.

Kiili, K., Moeller, K., & Ninaus, M. (2018). Evaluating the effectiveness of a game-based rational number training—in-game metrics as learning indicators. *Computers & Education*, *120*, 13–28.

Klein, G., Phillips, J. K., Rall, E. L., & Peluso, D. A. (2007). A data-frame theory of sensemaking. In R. R. Hoffman (Ed.), *Expertise out of Context: Proceedings of the Sixth International Conference on Naturalistic Decision Making* (pp. 113–155). Lawrence Erlbaum Associates.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253.

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessments in Education*, *2*(8), 1–24.

Levy, R. (2019). Dynamic Bayesian network modeling of game-based diagnostic assessments. *Multivariate Behavioral Research*, *54*(6), 771–794.

Levy, R. (2020). Implications of considering response process data for greater and lesser psychometrics. *Educational Assessment*, *25*(3), 218–235.

Lindner, M. A., & Greiff, S. (2023). Process data in computer-based assessment: Challenges and opportunities in opening the black box. *European Journal of Psychological Assessment*, *39*(4), 241–251.

Liu, T., & Israel, M. (2022). Uncovering students' problem-solving processes in game-based learning environments. *Computers & Education*, *182*, 104462.

Mayer, R. E. (2019). Computer games in education. *Annual Review of Psychology*, *70*(1), 531–549.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *60*(1), 577–605.

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

Mislevy, R., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining*, *4*(1), 11–48.

Mislevy, R., 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.

Mislevy, R., Oranje, A., Bauer, M. I., von Davier, A. A., Hao, J., Corrigan, S., Hoffman, E., DiCerbo, K., & John, M. (2014). Psychometric considerations in game-based assessment. White paper, GlassLab Research, Institute of Play.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.

Pellegrino, J. W., & Quellmalz, E. S. (2010). Perspectives on the integration of technology and assessment. *Journal of Research on Technology in Education*, *43*(2), 119–134.

Petri, G., & Gresse von Wangenheim, C. (2017). How games for computing education are evaluated? A systematic literature review. *Computers & Education*, *107*, 68–90.

Pirolli, P., & Russell, D. (2011). Introduction to this special issue on sensemaking. *Human-Computer Interaction*, *26*(1), 1–8.

Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist*, *50*(4), 258–283.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage.

Raykov, T., & Marcoulides, G. A. (Eds.). (2006). *A first course in structural equation modeling*. Lawrence Erlbaum Associates.

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer.

Reese, D. D., Tabachnick, B. G., & Kosko, R. E. (2015). Video game learning dynamics: Actionable measures of multidimensional learning trajectories. *British Journal of Educational Technology*, *46*(1), 98–122.

Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer*. Sage.

Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. *Computers in Human Behavior*, *116*, 106647.

Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. *Psychometrika*, *85*(2), 378–397.

Tenorio Delgado, M., Arango Uribe, P., Aparicio Alonso, A., & Rosas Díaz, R. (2016). TENI: A comprehensive battery for cognitive assessment based on games and technology. *Child Neuropsychology*, *22*(3), 276–291.

Tokac, U., Novak, E., & Thompson, C. G. (2019). Effects of game-based learning on students' mathematics achievement: A meta-analysis. *Journal of Computer Assisted Learning*, *35*(3), 407–420.

van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics*, *28*, 369–386.

Vendlinski, T. P., Delacruz, G. C., Buschang, R. E., Chung, G. K. W. K., & Baker, E. L. (2010). Developing high-quality assessments that align with instructional video games. CRESST Report 774, University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science*, *16*(4), 409–421.

Weiner, E. J., & Sanchez, D. R. (2020). Cognitive ability in virtual reality: Validity evidence for VR game-based assessments. *International Journal of Selection and Assessment*, *28*(3), 215–235.

What Works Clearinghouse. (2015). WWC review of the report "The Effects of Math Video Games on Learning."

Wouters, P., Nimwegen, C., Oostendorp, H., & Spek, E. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology*, *105*, 249.

Wüstenberg, S., Greiff, S., & Funke, J. (2012). Complex problem solving—more than reasoning? *Intelligence*, *40*(1), 1–14.

Xiao, Y., Veldkamp, B., & Liu, H. (2022). Combining process information and item response modeling to estimate problem-solving ability. *Educational Measurement: Issues and Practice*, *41*(2), 36–54.

Zhang, S., Wang, Z., Qi, J., Liu, J., & Ying, Z. (2023). Accurate assessment via process data. *Psychometrika*, *88*(1), 76–97.

Zhu, S., Guo, Q., & Yang, H. H. (2023). Beyond the traditional: A systematic review of digital game-based assessment for students' knowledge, skills, and affections. *Sustainability*, *15*(5), 4693.

Zumbo, B. D., Maddox, B., & Care, N. M. (2023). Process and product in computer-based assessments. *European Journal of Psychological Assessment*, *39*(4), 252–262.

Authors

TIANYING FENG is a doctoral student in the Education - Advanced Quantitative Methodology program at UCLA and a research assistant at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), SEIS Building, Los Angeles, CA 90095-1522; tfeng0315@ucla.edu. Her primary research interests include technology-based measurement and learning research and statistical computing.

LI CAI is a Professor of Education in the Advanced Quantitative Methodology program at UCLA and Director of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST), 315 SEIS Building, Los Angeles, CA 90095-1522; cai@cresst.org. His primary research interests include psychometrics and statistical computing.