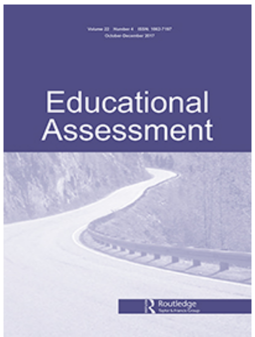


Educational Assessment

ISSN: 1062-7197 (Print) 1532-6977 (Online) Journal homepage: http://www.tandfonline.com/loi/heda20

Modeling Student Cognition in Digital and Nondigital Assessment Environments

Kristen E. DiCerbo, Yuning Xu, Roy Levy, Emily Lai & Laura Holland

To cite this article: Kristen E. DiCerbo, Yuning Xu, Roy Levy, Emily Lai & Laura Holland (2017) Modeling Student Cognition in Digital and Nondigital Assessment Environments, Educational Assessment, 22:4, 275-297, DOI: 10.1080/10627197.2017.1382343

To link to this article: https://doi.org/10.1080/10627197.2017.1382343

Published online: 13 Oct 2017.

EDUCATIONAL ASSESSMENT
2017, VOL. 22, NO. 4, 275–297
https://doi.org/10.1080/10627197.2017.1382343

Modeling Student Cognition in Digital and Nondigital Assessment Environments

Kristen E. DiCerbo[a], Yuning Xu[b], Roy Levy[c], Emily Lai[a], and Laura Holland[a]

[a]Pearson, London, United Kingdom; [b]SRI International, Menlo Park, California; [c]Arizona State University

ABSTRACT
Inferences about student knowledge, skills, and attributes based on digital activity still largely come from whether students ultimately get a correct result or not. However, the ability to collect activity stream data as individuals interact with digital environments provides information about students' processes as they progress through learning activities. These data have the potential to yield information about student cognition if methods can be developed to identify and aggregate evidence from diverse data sources. This work demonstrates how data from multiple carefully designed activities aligned to a learning progression can be used to support inferences about students' levels of understanding of the geometric measurement of area. The article demonstrates evidence identification and aggregation of activity stream data from two different digital activities, responses to traditional assessment items, and ratings based on observation of in-person non-digital activity aligned to a common learning progression using a Bayesian Network approach.

The digital revolution has brought about sweeping changes to most aspects of modern life. Digital computing and communication technology have dramatically changed how we gather and consume information and how we interact with others in almost every aspect of our lives (Mayer-Schonberger & Cukier, 2013). In addition, digital activity has led to collection and storage of information about individual activity at a scale never before seen (DiCerbo & Behrens, 2014). These capabilities should fundamentally change how we think about education in general (Collins & Halverson, 2009), and assessment in particular (DiCerbo & Behrens, 2014). Existing methods of understanding what students know and can do were developed in response to constraints around how little of the learning process could be observed by anyone outside the classroom (DiCerbo & Behrens, 2012). These constraints are now eased; data from student interactions in digital learning environments can be captured "invisibly," or without conspicuous, isolated assessment events. With this new data, however, come new challenges in domain definition, activity design, evidence identification, evidence accumulation, and communication of results.

We can take advantage of our ability to capture and store information about individual interactions and activities in education to increase our ability to understand what students know and can do. For example, if we are interested in creating a model of algebraic thinking, we might have a specific game in which players are planting a garden and must make a variety of decisions based on mathematical thinking and fundamental algebraic ideas. In the design of this game, the authors specify actions that differentiate experts and novices in this area and create tasks that elicit those actions, for example, computing the number of bricks for a raised bed. The system records the actions and time players take in the game, resulting in a log file. Some of the elements of the log file

CONTACT Kristen E. DiCerbo kristen.dicerbo@pearson.com Pearson, 3075 W Ray Rd., Chandler, AZ 85226.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/HEDA.

are identified and combined as evidence of players' skill level in algebraic reasoning. In traditional assessment, inferences about student knowledge, skills, and attributes still largely come from whether students ultimately get a correct result or not. Digital activity allows for tracking the process students use to solve a problem, not just their final product. This process data can be used to provide additional evidence about student proficiency. However, in order for this sequence from activity to evidence to inference to happen, and result in a valid measure of a player's reasoning, the tasks, evidence, and models must be carefully designed and tuned.

In addition, the scenario described above results in estimates of proficiency for a small number of skills from one source. We know that skills may or may not generalize across environments, and we know that evidence from individual performance tasks may or may not generalize well to other tasks (Keller, Clauser, & Swanson, 2010). Ideally, we would combine information collected across multiple environments where students are performing a particular skill. In our example of algebraic reasoning, we might have the game, plus an online adaptive homework environment, plus reference videos students can watch. We may want to combine this information from multiple contexts, including time variables, correctness variables, and motivation indicators, to make inferences about algebraic reasoning. Shute, Leighton, Jang, and Chu (2016) described such a vision as the future of assessment. They imagined a future in which students would progress through the year engaged in different learning contexts, all of which capture and measure growth in valuable cognitive and noncognitive skills. These authors noted that this will require high-quality, ongoing, unobtrusive assessments embedded in various technology-rich environments that can be aggregated to inform a student's evolving competency levels at various grain sizes. In order to do this, we gain by having clear models of the domain, how rich performances may constitute evidence about skills, and statistical methods for combining this disparate information.

Evidence-Centered Design (ECD; Mislevy, Steinberg, & Almond, 2003) provides a framework for conceptualizing this assessment activity. ECD encourages the development of an assessment by articulating the claims we wish to make with an assessment, the behaviors that would provide evidence for those claims, and the tasks or activities that would elicit those behaviors. Assessment design requires the specification of (a) a student model, identifying the constructs of interest; (b) a task model, specifying the activity to be performed; and (c) an evidence model to link the two. The evidence model consists of two parts: evidence identification, in which performances are evaluated to yield observed summaries, and the measurement model that connects these observables to the student model.

Learning progressions (LPs) provide one way to ground a student model. LPs lay out the major steps learners take as they move toward mastery of a given domain (Corcoran, Mosher, & Rogat, 2009). They are often expressed as a set of levels or stages of a particular skill as a learner progresses from less to more sophisticated understandings. The steps of progressions are not merely a description of a table of contents of a textbook or a scope and sequence of a curriculum. Rather, they are based on empirical research on student learning and focus on the large conceptual changes that occur in the move from novice to expert, which may be accomplished through multiple means.

A clearly specified LP can then allow for the creation of activities that align to specific stages. From a learning perspective, activities can be designed to specifically move learners from one stage to another. From an assessment perspective, activities can be designed to elicit evidence about stages a student has mastered and stages on which they are working. The evidence model then becomes the means by which individual pieces of evidence are extracted and scored from the students' performance and how those individual pieces are combined. It should be noted that an individual student's performances with respect to levels of LPs can show striking variability in different contexts and different content areas (Sikorski & Hammer, 2010). Different pieces of evidence related to the same stage may provide conflicting information, making thoughtful combination of evidence important.

In classrooms, scores across multiple assignments are typically simply averaged to get an overall grade. Teachers either assign weights or differential totals to assignments to signify which ones are more important, which could mean better indicators of overall skill with the topic of the course, but might also include things like the size and time requirements of the assignment. This weighted average method can be problematic for two reasons. First, new forms of assessment provide evidence that is not of the form: "number correct out of total." Things like latency and sequence cannot be captured in percentages, and are not easily translated to such metrics. Second, the knowledge, skills, and abilities of interest are latent. Percentages summarize how someone performed on particular tasks (taken with a particular context), and are often used to make inferences about proficiency in the domain broadly conceived, but performance on a given task is not a perfect measure of proficiency. People can "get lucky" and guess correctly on multiple choice items, resulting in higher scores than that which would reflect their proficiency. They can get sick, fatigued, or distracted and score lower than that which would reflect their proficiency.

Modern measurement models are constructed by setting up observable variables as stochastically dependent on latent student model variables. This in turn supports the use of the evidence from observed performance to reason about someone's "likely" or "probable" skill level, using probabilities to model the uncertainty in the inference from the performance to beliefs about the student's proficiency (Levy & Mislevy, 2016). Aggregating pieces of evidence collected across single, non-traditional assessment activities, such as a game or simulation, in order to make inferences about learners' mastery of stages of an LP has been demonstrated in a number of instances (Mislevy et al., 2014; West et al., 2010). However, the challenges of combining information across multiple digital activities compound those we see in aggregation from a single activity. For example, the well-known issue of testlets in traditional assessment (Wainer, Bradlow, & Wang, 2007) tells us that evidence that comes from one source is likely to share more variance than completely independent evidence. Ignoring testlet effects may render observables conditionally dependent, which can lead to incorrect estimates of the values of parameters and variables, as well as the precision of those estimates, yielding situations in which we overstate our precision and overestimate reliability (Bradlow, Wainer, & Wang, 1999; Sireci, Thissen, & Wainer, 1991). This begs the question: how then should we account for the context effects of evidence that comes from a game versus evidence that comes from another digital activity, for example?

This article demonstrates an approach to evidence identification and aggregation of activity stream data from multiple sources. The evidence comes from two different digital environments, responses to traditional assessment items, and ratings based on observations of in-person, non-digital activity, all aligned to make inferences about students in terms of mastery of stages on a common LP. The worked example in this article first demonstrates an approach to modeling different relationships between stages in a LP (pre-requisite versus independent stages). Second, it models the inclusion of evidence that can inform beliefs about mastery for multiple stages. Third, it includes evidence based on both the processes students use to solve problems and their final responses. Finally, it combines evidence from digital and non-digital, and traditional and nontraditional assessments. In doing so, it provides an exemplar for how to incorporate the influence of context effects in a model.

There are a variety of potential models that could be used to model cognition from evidence gathered from assessment (Rupp & Leighton, 2016). To conduct this work, we employ Bayesian networks (BNs; Almond, Mislevy, Steinberg, Yan, & Williamson, 2015), which offer several advantages. First, as a latent variable model, we accrue the benefits of modern measurement modeling discussed above. Second, latent proficiency variables in BNs are discrete, which aligns with the stage-based notions of proficiency in LPs, and supports fast computing of posterior distributions to make inferences about students. Central to the goals of this work, we believe a BN more easily allows researchers to account for evidence that relates to multiple latent variables, and address evidence that does not conform to traditional "correct/incorrect" judgements. In addition, BNs allow for researchers to specify relationships among latent variables, and model learning when present (Reye, 2004). BNs have been used in research on LPs (West et al., 2010), and studies involving context effects when multiple observations are obtained from the same source of evidence (Almond, Mulder, Hemat, & Yan, 2009; Levy & Mislevy, 2004).

This work demonstrates the suitability of latent variable models for modeling performances on these tasks. In particular, it demonstrates the utility of these models for synthesizing evidence from across the various tasks, supporting an interpretation of latent variables in terms of proficiency, all the while acknowledging the variation across the contexts that provide that evidence. In summary, this project uses evidence from multiple sources to make inferences about learners' mastery of stages on a LP.

This work is necessary if we are to move toward a world where we are able to provide a system of assessment in which we are able to provide more information about student knowledge and skills on an on-going basis to students and teachers with fewer traditional tests. It provides the foundation for moving towards the vision of the future of assessment in which teachers and students can take advantage of the explosion of digital data to make better decisions while completing fewer activities that look like traditional tests. In order for this to become a reality, we need to develop and examine methods such as those described in this article.

Methods

Participants

Data from 131 third grade students in six U.S. classrooms in Indiana were analyzed in the study. Third grade students were used because the knowledge and skills targeted here are part of the third grade Common Core standards. We were confident that students of this age would be familiar with digital game interactions given that research suggests 91% of students in this age group play video games (NPD Group, 2011). Approximately 78% of students at the school were Caucasian, 5% Black, and 5% Asian. Approximately 2% of students were identified as Hispanic. 17% of students were eligible for free or reduced lunch. Overall, the school scores in the top 20% of Indiana elementary schools on the Indiana statewide achievement test (ISTEP).

Materials

The *Insight Learning System* is a series of digital and non-digital activities aligned to an LP, with levels from lesser to greater sophistication in understanding of geometric measurement of area. The goal of the Insight Learning System is to investigate not whether students have memorized the formula for area, but rather understand conceptually what that formula means and that the resulting number represents the number of unit squares that fill a space. The purpose of the measurement here is, broadly speaking, somewhat agnostic between formative vs. summative, and are low-stakes as opposed to high-stakes. The modeled proficiencies are fairly fine-grained from the perspective of the LP, in the sense that we are drilling down to focus on a quite specific part of the larger LP. This grain-size is much finer than many summative assessments, though we acknowledge that even finer-grain sizes are possible.

To build the LP, we consulted mathematics education research literature, focusing on research on the development of student understanding of concepts and ideas related to geometric measurement of area. We synthesized relevant resources from a number of researchers (Barrett et al., 2011; Battista, Clements, Arnoff, Battista, & Borrow, 1998; Sarama & Clements, 2009). We also relied on the work of the Common Core Standards Writing Team, which has produced several draft progressions that attempt to tie existing learning sciences research to specific Common Core State Standards in order to lay out a hypothetical progression of topics. Using the draft progression as a foundational document, we consulted original research from mathematics education to extract and define a series of distinct stages through which students might pass on their way to learning geometric measurement of area. To give a high-level view of the context of this work, Figure 1 is a graphical representation of the LP, demonstrating how it ranges from kindergarten through grade 4 and contains separate strands for length, area, figure composition and decomposition, and geometric shapes. For each stage, we created a description of the stage, defined a set of observable behaviors that could serve as evidence for that stage, and identified potential errors or

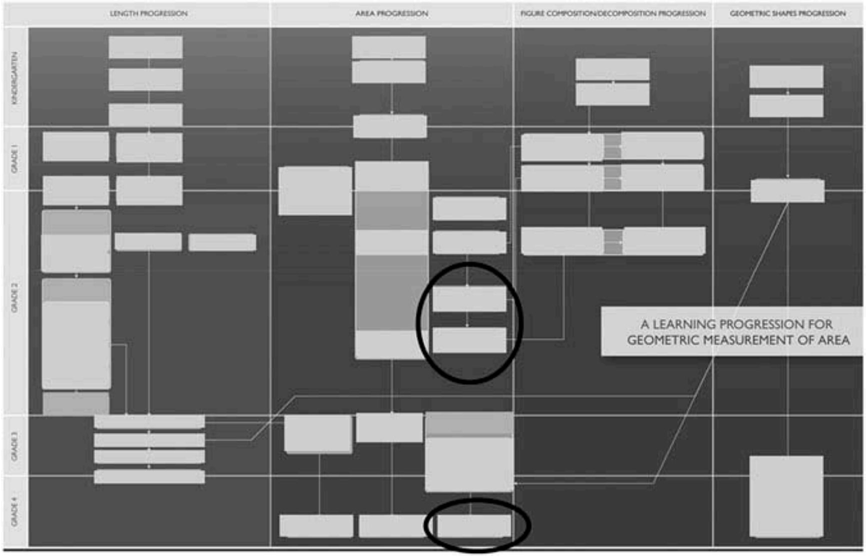

LENGTH PROGRESSION
AREA PROGRESSION
FIGURE COMPOSITION/DECOMPOSITION PROGRESSION
GEOMETRIC SHAPES PROGRESSION
KINDERGARTEN
GRADE 1
GRADE 2
GRADE 3
GRADE 4
A LEARNING PROGRESSION FOR GEOMETRIC MEASUREMENT OF AREA

Figure 1. Schematic of learning progression for geometric area. The circled squares are those investigated in this article.

misconceptions where possible. Further details on the broader LP represented in Figure 1 can be found in Lai, Kobrin, Nichols, and Holland (2015). In this article, we focus on a subset of stages (circled in Figure 1) concerning the concept of area units that appear to lead up to students mastering the area formula (i.e., “length x width”). At the stage *Using area units to measure area* (the top square circled in Figure 1) students perceive 2D shapes as collections of single area units and use those units to reason about area. Students at this stage can demonstrate their understanding by counting individual area units to determine the area of a 2D shape. Specific classroom behaviors may include pointing to each unit with fingers as they count, writing their counts on the area units, or drawing a check mark on each unit as they count. In the game context, a student may choose to manipulate only single units, creating a large number of them and arranging them individually without using more advanced tools to combine them.

In the stage *Using area composites to measure area* (the middle square circled in Figure 1), students perceive 2D shapes as collections of area composites and use two-level composites to reason about area. That is, instead of using single units alone, students use single units to make a composite unit, such as a column or row. This stage represents a higher level of cognitive sophistication than using single units to measure area. They may then either repeatedly add (e.g., 5 + 5 + 5) or multiply (e.g., 5 × 3) to find the area. Working with technology, a student may create single units, adhere the units into either columns or rows, then iterate the rows or columns- rather than the individual units- to fill the space. We hypothesize that *Using area units* is a prerequisite stage to *Using area composites*.

A third hypothesized stage is *formula proficiency* (the bottom square circled in Figure 1), requiring students to retrieve and apply the area formula. Students at this stage readily recognize 2D space problems as area problems and retrieve the familiar “length x width” area formula to solve the problem. Misconceptions at this stage include multiplying all or any labelled side lengths, confusing area with perimeter, applying the rectangular area formula to non-rectangles, and multiplying opposite rather than adjacent sides. The two area unit stages are not necessarily prerequisites for the formula stage. We hypothesize that it is possible to retrieve and apply the formula in a rote manner without having an understanding of area units. Indeed, math educators lament that in many classrooms, the concept of area is taught by having students memorize the area formula without connecting it to the structure of rectangular arrangement of unit squares (Baturo & Nason, 1996; Simon & Blume, 1994).

We designed a collection of learning and assessment activities aligned to these stages, including a digital game, a set of online performance tasks, a set of paper-based classroom activities, and a set of traditional tests. The information we gleaned from our review of the learning sciences research directly informed the design of these activities. In particular, we set out to observe the kinds of verbal and nonverbal behaviors identified in the literature as evidence of student understanding by strategically manipulating those assessment features shown to elicit that evidence. The dataset included activity stream data from the game and digital performance tasks, human ratings of mastery based on a paper-and-pencil classroom activity, and responses to traditional assessment items.

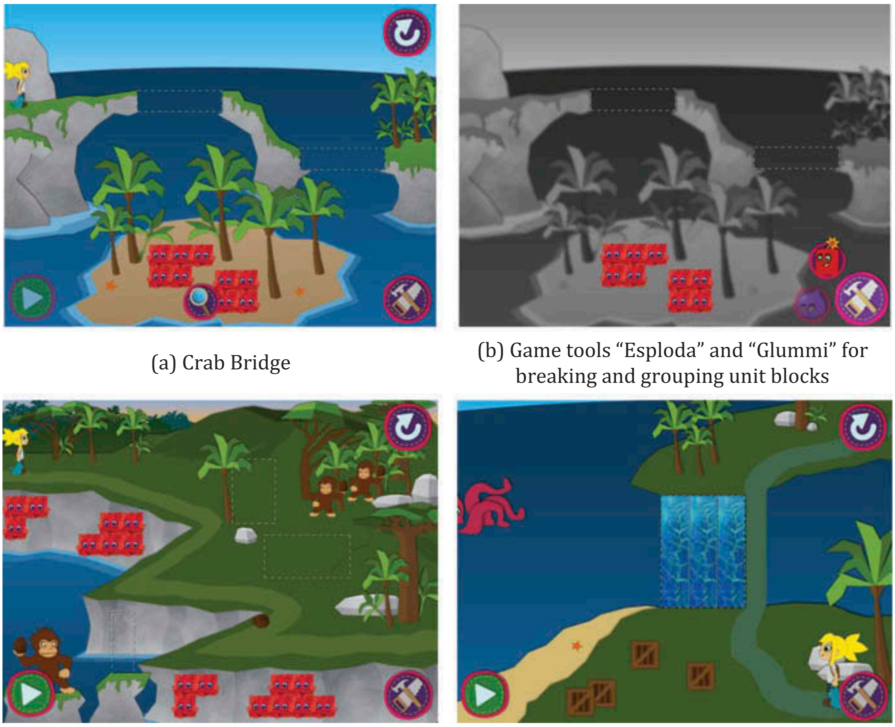
Game

The *Alice in Arealand* game (DiCerbo, Crowell, & John, 2015) was created using principles of Evidence-Centered game Design (Mislevy et al., 2014) with multiple iterations to ensure game actions were eliciting intended evidence to support inferences about mastery of LP stages. Players in the game help their new friend Alice and her pal Flat Cat navigate through a 2D world resolving challenges from things like a yeti and a kraken by demonstrating mastery of various stages related to manipulation of area units. The game has characters that function as tools to glue units together (character name Gluumi), break them apart (character name Esploda), and make copies of single or composite units (character name Multi). The early game levels focus on earlier stages of the LP not described here. The levels that provide evidence pieces relevant here are called Crab Bridges, Evil Monkey, and You Kraken Me Up!

In Crab Bridges, players must build a bridge by filling in a defined space with unit squares. As shown in Figure 2a, the unit squares look like crabs, and players must break up groups of crabs (composites) and reattach the correct number to fill the designated areas. Figure 2b shows the two tools, Esploda and Gluumi, for breaking and re-grouping the unit blocks in the lower right corner. The most efficient strategy involves using the Esploda tool twice to break up two groups of crabs. It was hypothesized that student who had mastered single area units and composites would use Esploda two or more times because they understand the need to get to the individual unit squares while nonmasters were expected use Esploda less frequently. In Evil Monkey (Figure 2c), players must build barriers to block monkeys' throwing coconuts at Alice and Flat Cat. To do so, they can either break up a number of existing composites or keep some of the already correctly shaped composites to fill the defined areas. It was hypothesized that students who have mastered the idea of composites would not break up the composites that are already of correct size, and thus would use the Esploda tool less frequently. Finally, in You Kraken Me Up! (Figure 2d), players must build a barrier so a kraken can't get to Alice and Flat Cat. They need twelve unit squares for the blockade but are only given four. Players can choose to make copies of the individual squares and then build the barrier or first make a composite from the four blocks and then make copies of the composite to fill the space. It was hypothesized that students who had mastered single units but not composites are more likely do the former while students who had mastered composites are more likely do the latter. Students could access hints on each level by clicking a magnifying glass. If their solution to a level was not correct, they were looped back to play the level again. Performance and actions of students' multiple attempts at each game level were modeled in this work.

Performance tasks

We designed four online performance tasks. Each task included 10–15 discrete activities or interactions and was designed to take students approximately 25 minutes to complete. Two of the tasks focused on the theme of designing a zoo and two around designing a playroom. For the purposes of this article, we focus on two activities that provided information about area unit proficiency and 13 activities that provided information about the use of the area formula. The two activities that provide information about area unit proficiency are both involved in designing the zoo, one involving a hippopotamus enclosure (Hippo) and the other a reptile house (Reptile). In both cases, students are presented with a rectangle with a grid of unit squares overlaid and asked to indicate how many



(a) Crab Bridge

(b) Game tools "Esploda" and "Glummi" for breaking and grouping unit blocks

(c) Evil Monkeys

(d) You kraken me up!

Figure 2. Screen shots of three game levels.

squares are required to fill the space. Side lengths of the shapes are not provided. Importantly, they are asked to show their work with a calculator from which we capture all of their entries. Students who have not mastered the idea of area units would be expected to get this question wrong. Students who have mastered area units but not composites would be expected to count the squares, and students who have mastered composites would be expected to use addition or multiplication to find the number. During initial testing, it was observed that students who counted the squares were likely to simply enter the result in the calculator with no formula. Students who used the formula generally entered it in the calculator, even if they were able to do the calculation in their head. We hypothesized that the use or not of a formula in the calculator could provide evidence of mastery or nonmastery of *Using area composites*.

The activities requiring the mastery of the area formula looked similar to many traditional assessment items assessing area, requiring students to apply the length times width formula to a rectangular shape where adjacent side lengths are labeled, but the internal unit structure of the shape is not visible. Evidence from these activities is correctness of the response.

Classroom activities

We also designed a series of paper and manipulative-based classroom activities intended to be implemented by teachers. After completing these activities, teachers were asked to enter their judgments of student mastery of particular stages. One of the classroom activities, called Flying with Squares, aligned to the *Using area units* stage of the LP. The activity required students to find the area of flying carpets for a number of figures, each demonstrating successively reduced scaffolding: first the entire carpet is covered in squares, then with a grid of dots, and finally with just a reference unit square outside the shape. Students who have mastered the *Using area units* stage should correctly find the area. However, the activity does not differentiate students who may also have mastered composites; they would also correctly find the area of the carpets. Therefore, this activity is hypothesized to provide evidence about one stage in the progression. For this article, due to constraints on teachers' time, ratings of mastery based on student work were completed by researchers reviewing student work products. Two researchers worked together to develop the evidence rules for mastery and non-mastery judgements. They worked together on practice samples (not part of this study) to achieve sufficient agreement. They then individually graded two work samples and then compared ratings. There was 89% agreement on the ratings. The raters discussed disagreements until they reached consensus in order to determine which judgement was used in the modeling. The key for this article was to develop ways to model expert judgment alongside digital information. Future models could be built that take into account the variability across teacher judgments.

Traditional tests

Finally, traditional assessments were constructed by collecting area and perimeter-related questions from items released from state assessments and curricula. Procedurally, there were two such assessments, each consisting of 20 questions, some of which were selected response questions and others constructed response questions (entering calculated area).

As detailed further in the next section, one of these assessments was administered prior to the aforementioned activities (i.e., the game, performance tasks, classroom activities), and the other was administered after those activities. Reflecting this temporal order, we refer to these tests as the pre-test and post-test, respectively. It should be noted that the post-test is not meant to be the "gold standard" or best measure of student proficiency against which other measures in the study are to be judged. Similarly, in this work, we do not view the situation as one in which we are examining the extent to which other activities (i.e., the game, performance tasks, classroom activities) contribute to or predict changes from the pre-test to post-test. Rather, the pre-test and post-test are viewed as two of many sources of evidence, which have in common a format, method of administration, and type of data collected. Our focus is on the proper way to model performance on these assessments in conjunction with the others described above. In particular, we focused on the 13 pre-test and 10 post-test items related to area. Of these, five pre-test items and two post-test items required counting unit squares to find the area of a shape, indicating mastery of *Using area units*. Three pre-test and four post-test items required use of the area formula to find the area of a shape, indicating mastery of formula proficiency. Finally, five pre-test and four post-test items could be solved using either counting or the area formula (we call these "mix" items).

Procedure

In all six classrooms (each taught by a different teacher), teachers implemented the Insight Learning System over a period of about two weeks. Teachers first participated in a series of face-to-face professional development sessions aimed at introducing them to the LP, the components of the system, and their intended use. Before beginning their instruction on area concepts, students completed a short pre-test composed of published or released items assessing area concepts. Students then completed the first two online performance tasks. Over the course of the next 6–7 days, during their regular math instruction, students alternated playing the digital game, participating in small group activities, engaging in whole class instruction, and completing the paper-based formative activities. At the conclusion of their instruction in area concepts, students completed the second set of two online performance tasks, and took the post-test composed of released area items.

The teachers did not provide any direct instruction and served largely as coaches when students got stuck activities. They did not provide instructional feedback on any of the activities. There were not differences in pretest-posttest scores across classrooms, suggesting the teachers did not engage in behaviors that differentially effected student learning across classes.

The brief amount of instruction provided over the course of the study was not sufficient to result in learning over the course of the study (Lai, Kobrin, & DiCerbo, 2016). After multiple imputation to address missing data, the pre-test mean number of questions correct was 3.81 and the post-test mean was 4.14 out of a total of 20 items (the means were even closer without imputation). This article focuses not on learning during the course of activities (e.g., from pre-test to post-test), nor on the effectiveness of the other activities (i.e., the game, performance tasks, classroom activities) conceived of as an intervention, but on using evidence from the various sources to develop assessment models.

Bayesian network model

A BN model was used in this study to aggregate evidence from the multiple assessment sources. A BN model in such contexts is often comprised of two interacting parts: the student model and the evidence model. The student model consists of latent variables that represent student learning proficiencies to be measured, along with directed links that represent known or hypothesized relations among proficiencies. The evidence model contains variables that correspond to observables derived from evaluating performances on the activities and directed links that express how the observables relate to the latent proficiency variables.

Building the student model

To build a BN model, we first specified the latent variables in the student model. As described, we focused on three particular stages of the LP: *using area units to measure area, using area composites to measure area*, and *formula proficiency*. Two latent proficiency variables were defined for the three stages: *Area Unit Proficiency* and *Formula Proficiency. Area Unit Proficiency* has three levels corresponding to stages of understanding area units: nonmaster, single units, and collection of composites. *Formula Proficiency* has two levels of whether or not students have mastered the formula of calculating area: nonmaster and master.

In addition to the two proficiency variables, two latent context variables were introduced to model the contextual effect for evidence from the same sources. For games and performance tasks, we defined Game Effect and Task Effect to account for the potential associations among the observables from the same digital activities over and above the proficiency variables. Each of the two context variables has two levels: lowered and elevated. The term *Effect* is taken as a generic term to describe the association due to context. The meanings of the context variables are of less concern in this study. That is, we might think of them as generic context variables (Levy & Mislevy, 2004) or as group factors (Reise, 2012). The two levels of each of the context variables are used to aid in characterizing students who, holding proficiency variables constant, perform worse in the environment (coded as being at the lowered level of the variable), as distinct from those who, holding proficiency variables constant, perform better in the environment (coded as being at the elevated level of the variable). Our primary aim for including the context variables is to help purify the proficiency variables when making inferences, as observables from the same assessment context tend to “cluster” together, in the sense of having shared variation.

All the four latent variables are modeled as a priori independent from each other in the network. Note that this does not mean that the variables will not be correlated in the final results. Relationships present in the data will emerge in the analysis.

Building the evidence model

Next we identified the specific features of the various activities as observed variables to build up the evidence model and linked these observables to the latent variables in the student model. In Figure 3, a graphic presentation of the complete BN model is presented. Latent variables are highlighted by shaded blocks with black bold frames in the figure, and observables from the same evidence sources

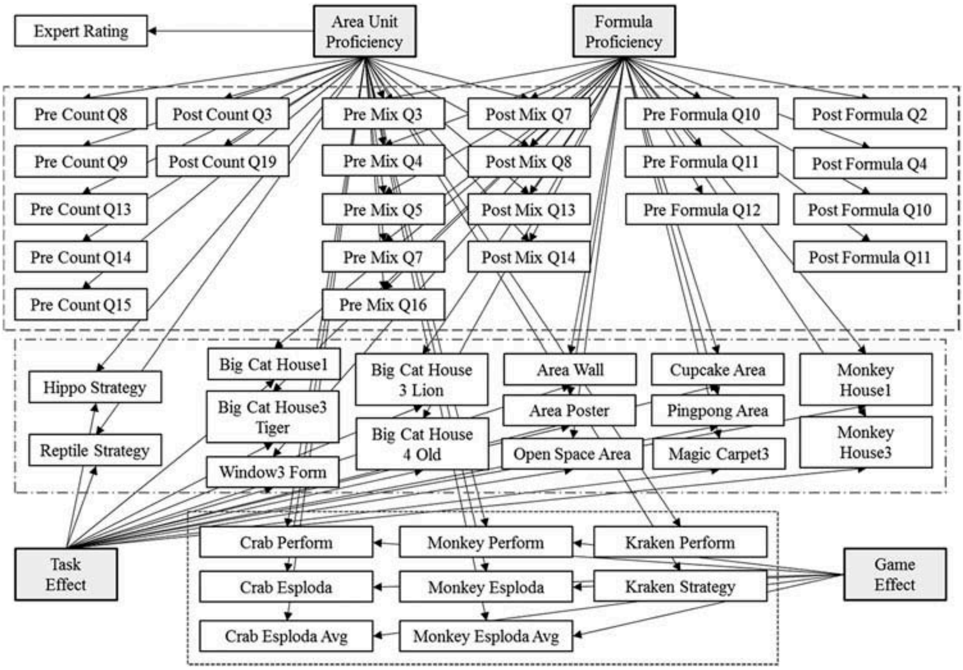


Figure 3. A graphical presentation of the BN model for modeling student learning proficiencies in the Insight Learning System. Dotted line = Observables from Game; dotted and dashed line = Observables from Performance Tasks; dashed line = Observables from Pre and Post Tests. "Expert Rating" is the observable from Classroom Activities. Shaded blocks represent latent variables.

are outlined with dotted lines. For the three game levels, log files were examined, and two types of variables were extracted: the completion level and the strategy. Game completion on each game level is quantified by the number of failure attempts before one's first success attempt. For Crab Bridges and Evil Monkey, game strategy is quantified by the overall frequency of using the explosion tool Esploda to break up composites during a game level, as well as the average frequency per attempt during that level. For You Kraken Me Up!, students' strategies are categorized into four groups. Two major strategies as discussed earlier are (a) multiplying individual blocks to the desired quantity and then gluing the copies to build the barrier (the Multi-Gluumi strategy) and (b) gluing individual blocks to a composite and then multiplying the composites to build the barrier (the Gluumi-Multi strategy). These two strategies are hypothesized to be highly correlated with the two advanced levels of *Area Unit Proficiency*. Using the Multi-Gluumi strategy was hypothesized to be indicative of the single units stage. Using the Gluumi-Multi strategy suggests students are thinking about composites of units, indicative of the collection of composites stage. Also, students can apply a mixed strategy of the previous two. Lastly, we identified some other strategies that are not necessarily more strongly associated with any of the three levels of *Area Unit Proficiency*. All the game observables are associated with *Area Unit Proficiency* (as opposed to *Formula Proficiency*).

Game observables from Crab Bridges and Evil Monkey are also associated with the context variable *Game Effect*, capturing that these observables arise from a common source, namely the student's interaction with the game. We initially modeled the observables from You Kraken Me Up! as dependent on this game effect context variable as well. Results from modeling it as such indicated that the conditional probabilities did not differ for different levels of the context variable (when holding the proficiency constant). Accordingly, in our final model, we did not include *Game Effect* as affecting the observables from You Kraken Me Up! Given that this is the penultimate level in the game, it may be that previous levels have taught and reinforced game tools and play, reducing or eliminating the effect of game context on the level.

In the performance tasks, the Hippo and Reptile activities are modeled as depending on *Area Unit Proficiency*, and the remaining 13 activities target *Formula Proficiency*. For Hippo and Reptile, we focused on the strategies students used to complete the activities instead of the correctness of their final answers. The strategies are summarized into four categories: multiply, add, blank (indicating students might calculate the area by counting), and other. For the remaining 13 performance task activities, correctness was modeled via binary observables. All performance task observables are associated with their purported proficiency variables and *Task Effect*.

Classroom activities are summarized by a single binary observable of expert rating, which evaluated whether or not students have mastered the single unit level of *Area Unit Proficiency*. Lastly, we included 23 pre- and post-test items, modeling 7 of the items as depending on *Area Unit Proficiency*, 7 on *Formula Proficiency*, and 9 on both. All the pre- and post-test items are binary variables and have no additional context variable associated with them, since each item featured a unique problem context.

To summarize, we developed a BN model with two latent learning proficiency variables, two latent context variables, and 47 observables associated with the four latent variables. Table 1 presents a condensed Q-matrix (e.g., Almond et al., 2015) specifying the links from the latent variables to the observables. For game and performance tasks, each of the observables is linked to one proficiency variable and one context variable. Such specification reflects our hypothesis about how evidence supports inferences about latent proficiencies in a complex assessment environment with multiple sources. The incorporation of latent context variables helps in refining inference making by accounting for the common variance among evidence from the same assessment source. The structure of the model reflected subject matter expert beliefs about the nature of proficiency, and how the observables derived from these activities constitute evidence about that proficiency. In addition, subject matter expert beliefs dictated a number of constraints on the model and prior specifications used in fitting the model, as described next.

Model fitting

We adopted a fully Bayesian approach to fitting the model, using WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2007). Discussions with subject matter experts yielded expressions of prior beliefs regarding the probabilities that govern the associations among the variables in the BN. These formed the bases of Dirichlet prior distributions for the unknown conditional probabilities (Almond et al., 2015). Generally the

Table 1. Q-matrix of specifying the links from latent variables to observables.

				Latent Variables			
Evidence Sources	Observables		Number of observables	Area Unit Proficiency	Formula Proficiency	Game Effect	Task Effect
Game	Completion Level	Crab Bridge	1	×		×	
		Monkey Evil	1	×		×	
		You Kraken Me Up!	1	×			
	Strategy	Crab Bridge	2	×		×	
		Monkey Evil	2	×		×	
		You Kraken Me Up!	1	×			
Performance Tasks	Strategy	Hippo	1	×			×
		Reptile	1	×			×
	Correctness	Other PTs	13		×		×
Traditional Tests	Correctness	Count Items	7	×			
		Formula Items	7		×		
		Mix Items	9	×	×		
Classroom Activities	Expert Rating		1	×			

hyperparameters of these prior distributions were specified with relatively minimal weight, akin to having 12 observations (Almond et al., 2015). This choice reflects a desire to have the results reflect a greater influence of the data collected from the students.

For the latent variables, the priors reflect subject matter experts' hypotheses about student distributions on each proficiency or context variable. In our hypotheses, the majority of the students were masters of the two proficiencies, given that the classes in the study had introduced the topics previously, and were familiar with the different assessment contexts, given the students in these classes were generally from demographic backgrounds with access to technology.

For the 47 observables, conditional probability tables (CPTs) were specified to embody our hypotheses about the probabilities of observing particular strategies or performances, conditioning on the relevant learning proficiency variables and the context variables. In our model, each observable is either associated with one or two latent variables. For those associated with two latent variables, the prior also reflects beliefs about the relative importance in how the two latent variables interact in influencing the observables. These predominantly followed hypotheses of compensatory relationships, where the low standing on one variable could be compensated by high standing on another (Almond et al., 2009).

In some cases, the priors reflected firm constraints designed to reflect substantive hypotheses, as well as to resolve the indeterminacies in discrete latent variables. Some of the constraints were specified initially, others were identified after fitting simpler models and considering their results. For example, for the pre- and post-test items that depended on *Area Unit Proficiency*, we constrained the conditional probabilities of correct to be the same for students at the single units and collection of composites levels, to reflect the idea that these items discriminate between nonmasters and students at either of these other two levels. For items that depended on both *Area Unit Proficiency* and *Formula Proficiency*, we constrained the conditional probabilities of correct for students at the collection of composites level of *Area Unit Proficiency* and mastery level of *Formula Proficiency* to be higher than the conditional probabilities for students at the single units level of *Area Unit Proficiency* and mastery level of *Formula Proficiency*. We also based the prior distributions for the probability of a correct response to the pre- and post-test items at the nonmastery levels of proficiency on what would be expected based on chance level of success from random guessing.

As an example of constraints from the game, using the Gluumi-Multi strategy (glue first, then copy) in Kraken was identified as being most indicative of being at the collection of composites level of *Area Unit Proficiency*. We constrained the conditional probability of using that strategy given the student was at the collection of composites level to be higher than if the student was at the single units level. Similarly, we hypothesized that use of the Multi-Glummi strategy (copy first, then glue) was most indicative of being at the single units level. Accordingly, we constrained the conditional probability of using that strategy given the student was at the single units level to be higher than if the student was at the collection of composites level. In such instances, we employed a gamma reparameterization for the Dirichlet distribution (Spiegelhalter et al., 2007) to enact such constraints.

The use of a probabilistic model to link evidence to stages of a LP is key to understanding how we can account for variability in student behavior. At any particular stage, students may exhibit behavior aligned with other different stages depending on a variety of contextual and person-specific factors. For example, students may elect to use a lower-level strategy at any given time, even though they are capable of using the higher-level strategy. The probabilities in the conditional probability tables allow us to account for this. They do not deterministically say that a student who has mastered a particular level will always use a particular strategy, nor that the use of a particular strategy is conclusive about what learning level the student has mastered. Rather, the model specifies a probability that a particular strategy will be used. When the model is used to make inferences about individuals, observing a student engaging in a certain behavior will adjust our belief about the probability of mastery, making it more or less likely, but not definite.

In using this model, we acknowledge that there may be further differences in the individuals (e.g., students capable of using a higher-level may vary in how likely they are to use a lower-level strategy) or contexts (e.g., activities may be differentially attractive to potential use of a lower-level strategy) that are not modeled. The probabilistic form of our model essentially aggregates over these finer-grained, unmodeled variations, and is expressed in a framework that allows for an extension of the model to address such variations should collateral information about them become available. Note that this idea of a probabilistic model aggregating over various situations is closely aligned with our principal goal of this work, namely to gather and aggregate evidence from multiple sources. Our view is that no one of these assessment activities (tasks, contexts) perfectly captures the proficiencies of interest, and, recalling that student performances can show variability in different contexts, we benefit by constructing a set of latent variables representing those proficiencies that draw their meaning from looking across these varied activities.

Model estimation

The complete BN model was fit to data in WinBUGS using three chains from dispersed starting values. A set of model constraints was imposed on the conditional probabilities to reflect substantive hypotheses, as well as to resolve the indeterminacy in latent group identification. Initially, we imposed equality constraints on probabilities of the pre- and post-test items to identify the two groups at the more advanced levels of *Area Unit Proficiency*. Additional constraints on the Hippo/Reptile and Kraken strategies emphasized our definitions of the single units and collection of composites levels of *Area Unit Proficiency* in terms of designing these performance tasks and game levels. The chains for the final BN model appeared to have converged within 500 iterations. To be conservative, we discarded the first 2,000 iterations for each chain and ran an additional 3,000 iterations, resulting in a total of 9,000 iterations for use in inference. The marginal posterior densities were unimodal and fairly symmetric, with departures from symmetry occurring when the densities for the parameters were located near a boundary of 0 or 1. We present numerical summaries of the posterior distributions of parameters in the BN in the next section.

Results

Data from 131 3rd grade students were analyzed in the study. Missingness on the 47 observables ranged from 3.82% to 58.78%. Observables with high missing rates included variables from You Kraken Me Up!, and post-test items 2–4, 10, 11, and 19. In terms of the game performance, all three game levels were found to have high completion levels. A high percentage of the students succeeded on the game levels on their first attempt (60.87% for Crab Bridge, 81.01% for Evil Monkey, and 47.27% for You Kraken Me Up!). On Crab Bridge, 26.96% of the students used the most efficient strategy (i.e., “2 times”). On Evil Monkeys, 27.85% of the students used the most efficient strategies (i.e., “2 times” or “3 times”). And on You Kraken Me Up!, 59.26% used the Multi-Gluumi strategy (copy first, then glue) and only 12.96% of them used the more efficient Gluumi-Multi Strategy (glue first, then copy). On the performance tasks, 59.52% and 76.42% of the students used the Multiply strategy on the Hippo (administered before game play and classroom instruction) and Reptile (administered after the game and classroom instruction) activities, respectively. The rest of the performance tasks had correctness rates ranging from 2.13% to 56.25%. Across the pre- and post-test items, the correctness rates ranged from 3.17% to 83.33%. Based on the classroom activity, 92.31% of the students were rated as at least on the single units level of *Area Unit Proficiency* by expert raters.

Results for the latent learning proficiencies and context effects

Summaries of the distributions of the latent variables from fitting the BN model are presented in Table 2. Most of the students (83.21%) were estimated at the most advanced level (i.e., collection of composites) of *Area Unit Proficiency*, only 4.66% of them were at the single units level, and 12.13% had not mastered the unit concept of area. Only about one third of the students were estimated to

Table 2. Posterior means of the probabilities for the latent variables in the BN model.

Area Unit Proficiency	Nonmaster	Single Units	Collection of Composites
	.12	*.05*	*.83*
Formula Proficiency	Nonmaster	Master	
	.64	*.36*	
Game Effect	Lowered	Raised	
	.14	*.86*	
Task Effect	Lowered	Raised	
	.17	*.83*	

have mastered *Formula Proficiency*. Less than 20% of the students were estimated to be at the low levels of the *Effect* variables, where their performance in these contexts suffered relative to their performance in other contexts.

Results for the observables

Next we describe how evidence from the various assessment environments are linked to the latent proficiencies by synthesizing the student data and our BN model. Table 3 presents posterior means of the CPTs for select observables, described in the following subsections.

Formula proficiency via pre-test and post-test items

Panel (a) presents CPTs for one pre-test item (Pre Formula Q10) and one post-test item (Post Formula Q10). Both the items targeted whether students have mastered *Formula Proficiency*. The posterior means indicated that students who have mastered the proficiency have much greater probability in obtaining correct answers on these two items.

Area unit proficiency via expert ratings

The results for Expert Rating, in panel (b), showed that the conditional probabilities of being rated as a master by experts are substantially higher for students at the two advanced levels of *Area Unit Proficiency* than for the nonmasters. Additionally, the two advanced proficiency levels are estimated to have fairly slight differences in terms of conditional probabilities. The implication is that Expert Rating assessed *Area Unit Proficiency*, but primarily differentiated the nonmasters of the proficiency from those who were either at the single units or collection of composites levels. It did not contribute much to differentiating between the single units and collection of composites levels.

Area unit proficiency via the game

The observable Kraken Strategy was hypothesized to differentiate the three levels of *Area Unit Proficiency* simultaneously (in contrast to Expert Rating, as just described). The results in panel (c) of Table 3 indicate that students at the nonmaster level tended to use the Multi-Gluumi strategy and a mixed strategy (probabilities of .36 and .35), and students at the single units level predominantly used the Multi-Gluumi strategy (probability of .66). These estimates are consistent with our prior hypotheses. On the other hand, students at the collection of composites level were expected to adopt the Gluumi-Multi strategy most often. However, the estimated results showed that using the Multi-Gluumi strategy was still the most probable response choice for these students. Importantly, the constraint that linked Gluumi-Multi with collection of composites and Multi-Gluumi with single units is reflected in the posterior. That is, the probability of using the Gluumi-Multi is higher for students at the collection of composites level (.26) than for students at the single units level (.10). The differences in the conditional probabilities between the two proficiency levels helped in separating students from one mastery level to the other. An observation of Multi-Gluumi serves to increase the probability that the student is at the single units level, while an observation of Gluumi-Multi serves to increase the probability that the student is at the collection of composites level.

Table 3. Posterior means of the conditional probability tables for observables in the BN model.

(a) Observables “Pre Formula Q10” and “Post Formula Q10” from the pre and post tests

	Pre Formula Q10		Post Formula Q10	
Formula Proficiency	Incorrect	Correct	Incorrect	Correct
Nonmaster	.94	.06	.86	.14
Master	.22	.78	.16	.84

(b) Observable “Expert Rating” from the classroom activities

	Expert Rating	
Area Unit Proficiency	Nonmaster	Master
Nonmaster	.59	.41
Single Units	.16	.84
Collection of Composites	.06	.94

(c) Observable “Kraken Strategy” from the game level You Kraken Me Up!

	Kraken Strategy			
Area Unit Proficiency	Other[a]	Multi-Glummi[b]	Mixed[c]	Glummi-Multi[d]
Nonmaster	.14	.36	.35	.15
Single Units	.10	.66	.14	.10
Collection of Composites	.10	.47	.17	.26

(d) Observable “Pre Mix Q3” from the traditional tests

	Formula Proficiency	Pre Mix Q3	
Area Unit Proficiency		Incorrect	Correct
Nonmaster	Nonmaster	.92	.08
Nonmaster	Master	.26	.74
Single Units	Nonmaster	.20	.80
Single Units	Master	.45	.55
Collection of Composites	Nonmaster	.77	.23
Collection of Composites	Master	.40	.60

(e) Observable “Crab Esploda” from the game level Crab Bridge

	Game Effect	Crab Esploda[e]			
Area Unit Proficiency		1- time	2 times	3 times	4+ times
Nonmaster	Lowered	.40	.09	.20	.31
Nonmaster	Raised	.27	.13	.28	.32
Single Units	Lowered	.09	.42	.30	.20
Single Units	Raised	.10	.42	.28	.20
Collection of Composites	Lowered	.05	.38	.30	.27
Collection of Composites	Raised	.02	.29	.25	.45

(f) Observable “Hippo Strategy” from the performance task Hippo

	Task Effect	Hippo Strategy			
Area Unit Proficiency		Other[f]	Blank[g]	Add[h]	Multiply[i]
Nonmaster	Lowered	.68	.11	.10	.11
Nonmaster	Raised	.71	.08	.13	.08
Single Units	Lowered	.32	.57	.11	< .01
Single Units	Raised	.31	.57	.11	< .01
Collection of Composites	Lowered	.21	.34	.08	.36
Collection of Composites	Raised	.17	.10	.07	.66

[a] Other: Strategies that do not belong to the other three strategies.

[b] Multi-Glummi: Students first made copies of the individual blocks and then glued the copies to build the barrier.

[c] Mixed: A strategy that is a combination of the Multi-Glummi strategy and the Glummi-Multi strategy.

[d] Glummi-Multi: Students first glued the four blocks into a composite and then made copies of the composite to build the barrier.

[e] Crab Esploda refers to the frequency of using the explosion tool “Esploda” to complete the Crab Bridge level. This observable characterizes the strategies students used on the game level.

[f] Other: Strategies that do not belong to the other three strategies

[g] Blank: No formula is entered, indicating that students might count the squares.

[h] Add: Students entered addition formula using the calculator.

[i] Multiply: Students entered multiplication formula using the calculator.

Formula proficiency and area unit proficiency via traditional items

Panel (d) presents a mix test item linked to both *Formula Proficiency* and *Area Unit Proficiency*. The item Pre Mix Q3 was hypothesized to have a high probability for correct if students are masters of *Formula Proficiency*, or if they are at least at the single units level of *Area Unit Proficiency*. Consistent with these beliefs, the posterior probabilities indicate that only when students are nonmasters of both proficiencies do they have a very low probability of answering the item correctly. What is more interesting for this item is that, at the single units level of *Area Unit Proficiency*, the probability of correct was reduced when students mastered *Formula Proficiency* as compared to students who did not. This interaction was reversed for the collection of composites level. This type of inhibitory relationship between two proficiency variables was not hypothesized in our priors but emerged in the data analysis.

Combining context effects in the game and performance tasks

Panels (e) and (f) of Table 3 display the posterior means of the conditional probabilities for two observables: Crab Esploda from the Crab Bridge game level and Hippo Strategy from the performance tasks. Both the observables are evidence for assessing different stages of *Area Unit Proficiency* and they are also affected by one context variable.

In panel (e), the Crab Esploda variable describes the strategy students used on this game level. The four levels of the observable are the different frequencies of using the explosion tool Esploda to break up crab composites in order to succeed the level. Students who had not mastered the area unit concept had very low probabilities of using the tool twice (i.e., “2 times”), which is the most efficient strategy. They were more likely to use the tool too seldom or too frequent. Students at the single units or collection of composites levels were more likely to use the most efficient strategy and were unlikely to underuse the tool. Underusing the tool may indicate players did not understand the need to get to individual unit squares. The context effects indicated by *Game Effect* appeared most strongly at the nonmaster and collection of composites levels of *Area Unit Proficiency*. At these two stages, students who are more negatively affected by the game context (i.e., likely at the lowered level of *Game Effect*) were estimated to use the Esploda tool less frequently than those who benefited more from the game format (i.e., likely at the raised level of *Game Effect*).

Lastly, panel (f) shows the results for Hippo Strategy, which was similar to that of Reptile Strategy (not shown). The focus on Hippo/Reptile Strategy was whether students at the single units versus the collection of composites levels used different strategies to calculate area. Consistent with our prior hypotheses, leaving a blank formula (indicating counting units) was found to be most probable for students at the single units level, and using the multiplication strategy had high probabilities for students at the collection of composites level. Moreover, the posterior reflected an inhibitory relationship between *Task Effect* and *Area Unit Proficiency* at the collection of composites level. Students likely negatively affected by the assessment context had a considerably lower probability of using the Multiply strategy. The same type of interaction appeared in most of the performance tasks.

Inferences regarding individual students

Once the BN was constructed, we could estimate the probabilities of the states of the latent variables for individual students from the observed evidence. In this section, we use examples of five individual students to illustrate how different patterns of observables impact the estimation of students’ latent proficiencies. The upper portion of Table 4 shows observed values for the five students while the lower portion shows the estimated probabilities of the levels of the latent variables.

Student 1 in the third column of Table 4 succeeded at the first attempt on the game level and performed a mixed strategy. In addition, the student did well on most of the performance tasks and the traditional test items. Given the consistent, strong performance across multiple assessments, student 1 was estimated as likely being at the highest levels of the latent variables: the collection of

Table 4. Observed data and estimated latent proficiencies for five students from sample data.

Observed Data						
Assessment		Student 1	Student 2	Student 3	Student 4	Student 5
Game	Kraken Performance	OneF_Suc[a]	Fail[b]	NA	Success[c]	MultiF_Suc[d]
	Kraken Strategy	Mixed[e]	Multi-Gluumi[f]	NA	Gluumi-Multi[g]	Multi-Gluumi[f]
Performance Tasks	Hippo Strategy	Multiply[h]	Other[i]	Blank[j]	Multiply[h]	Multiply[h]
	Reptile Strategy	Multiply[h]	Other[i]	Blank[j]	Multiply[h]	Multiply[h]
	Other PTs	12	1	0	4	2
Pre and Post Tests	Count Items	7	1	3	3	6
	Formula Items	6	0	0	3	1
	Mixed Items	8	1	4	4	5
Expert Rating		NA	Nonmaster	NA	Master	Nonmaster
Posterior Distributions and Estimated Group Membership of the Latent Variables[k]						
Latent Variables		Student 1	Student 2	Student 3	Student 4	Student 5
Area Unit Proficiency *(nonmaster, single units, composites)*		(<.01, <.01, **>.99**)	(**>.99**, <.01, <.01)	(<.01, **.61**, .39)	(<.01, <.01, **>.99**)	(<.01, <.01, **>.99**)
Formula Proficiency *(nonmaster, master)*		(<.01, **>.99**)	(**>.99**, <.01)	(**>.99**, <.01)	(**>.99**, <.01)	(**>.99**, <.01)
Game Effect *(lowered, raised)*		(<.01, **>.99**)	(.07, **.93**)	(.28, **.72**)	(.23, **.77**)	(**.59**, .41)
Task Effect *(lowered, raised)*		(<.01, **>.99**)	(.06, **.94**)	(.33, **.67**)	(.01, **.99**)	(.02, **.98**)

[a] OneF_Suc: Students failed one time before they succeed on the game level.
[b] Fail: Students failed on the game level.
[c] Success: Students succeed at their first attempt on the game level.
[d] MultiF_Suc: Students failed multiple times before they succeed on the game level.
[e] Mixed: A strategy that is a combination of the Multi-Glummi strategy and the Glummi-Multi strategy.
[f] Multi-Glummi: Students first made copies of the individual blocks and then glued the copies to build the barrier for the You Kraken Me Up! game level.
[g] Glummi-Multi: Students first glued the four blocks into a composite and then made copies of the composite to build the barrier for the You Kraken Me Up! game level.
[h] Multiply: Students entered multiplication formula using the calculator for the Hippo and Reptile tasks.
[i] Other: Strategies that do not belong to the other three strategies for the Hippo and Reptile tasks.
[j] Blank: No formula is entered, indicating that students might count the squares for the Hippo and Reptile tasks
[k] Numbers in parentheses below are estimated probabilities of being at each levels of the latent variables. For each latent variable, their specific levels are specified in the parentheses following the variables names.

composites level of *Area Unit Proficiency*, the mastery level of *Formula Proficiency*, as well as likely having higher scores due to the game context, all else being equal. Student 2, in contrast, failed at the game level and applied less efficient strategies on the game level. Student 2's performance on the performance tasks and traditional items was much worse than student 1. Based on expert rating, s/he was rated as a nonmaster of *Area Unit Proficiency*. As a result, student 2 had a greater than .99 probability of being a nonmaster of *Area Unit Proficiency* and a nonmaster of *Formula Proficiency*, as shown in the lower half of Table 4.

Student 3 had limited data on the observables. Compared to student 2, this student adopted the counting strategy (as indicated by "Blank") used for the Hippo and Reptile tasks and had better performance on the count and mixed items of the traditional tests. The better performance increased the probability of Student 3 being at more advanced levels of *Area Unit Proficiency*. Student 3 was estimated to have a .61 probability of mastery of single area units and a probability of .39 of mastery of composites.

Student 4, in comparison to students 2 and 3, was successful on the first try of the Kraken level and used the most efficient strategy. In addition, s/he used the Multiply strategy on the performance tasks and was rated as having mastered single area units by the teacher. This student had a probability greater than .99 of having mastered the composites stage. Lastly, student 5 had multiple failures before succeeding on the game level. In addition, the adopted strategy was the less efficient

Multi-Gluumi rather than Gluumi-Multi. The expert rating indicated non-mastery of single area units. However, additional information from the performance tasks shows that student 5 was able to apply the Multiply strategy, which is the evidence of being at the collection of composites level. In other words, this was a situation with conflicting evidence from different sources: The performance tasks provide evidence of mastery at the collection of composites level, while the game and expert rating provide evidence that mastery falls short of this level. The inclusion of the latent context variable *Game Effect* offers an alternative explanation to reconcile this pattern, namely that the student is likely at the collection of composites level, but their performance suffered in the game due to unique features of that context. The use of the BN with multiple latent variables representing proficiency and contextual effects offers a more nuanced way to manage and make sense of possibly conflicting evidence from different sources than traditional approaches.

Summary

To summarize, under our current model and given the data, it can be concluded that most of the students likely mastered the area unit conception at the collection of composites level, but more than half of them likely had not mastered the area formula. Traditional assessment formats such as standardized tests and ratings from teachers and professionals contributed to inferences about whether or not students mastered particular levels of the LP in a general sense. More importantly, evidence from particular digital activities such as games and online performance tasks allowed us to further investigate specific levels of learning proficiencies. This evidence provided detailed data on students' thinking processes while resolving tasks. In our case, strategies used on You Kraken Me Up! and on the Hippo and Reptile tasks were three key evidentiary pieces for differentiating students at the single units and collection of composites levels of area unit understanding.

The inclusion of latent context variables makes it possible to have alternative interpretations regarding the additional associations among observables from the same types of assessment. Our results indicated that unique assessment environments do make a difference in student performance on some particular levels of the performance tasks and games. The impact of the assessment contexts or specific features of the contexts might inhibit student performance on digital activities under some circumstances.

Discussion

Modeling the learning progression

A key goal of this work was to link evidence from non-traditional assessments to stages in an LP. As described above, we first created a student model with two variables: *Area Unit Proficiency* and *Formula Proficiency*. Note that these two variables incorporated three stages. Two of the stages, single units and collection of composites, were hypothesized to have a prerequisite relationship. We posit that students need to master the idea of single units before they master the idea of a collection of composites. Therefore, those stages were modeled as levels of a single latent variable. Understanding and application of the formula for area was hypothesized to be a separate variable. One of the issues raised by math educators is that students can learn to apply the formula as a rote algorithm without understanding the underlying logic of the area units. Therefore, this was modeled as a second latent variable.

The decisions about which pieces of evidence to link to the latent variables came largely as a result of the up-front design work done under the framework of ECD. That is, the tasks were specifically designed to elicit evidence about particular stages. You Kraken Me Up! was specifically designed to differentiate between single units and collection of composites stages. Note that tasks designed to indicate mastery of single units usually differentiated between nonmastery of single units and mastery of single units, but not between single units and collection of composites. The BNs allow the flexibility to model both the relationships between the stages and how the evidence does or does not differentiate between them.

The modeling is also able to account for tasks that could potentially provide evidence for multiple stages that are independent with each other. There were a number of items on the traditional pre- and post-tests that could potentially provide evidence for either the *Area Unit Proficiency* or the *Formula Proficiency* variables. In part, this is because the only piece of information recorded for these items is the answer selected; there is no information about how that answer was obtained. Most of these items involved the presentation of a rectangular grid of squares with the length and width noted on the sides, and the request to find the area of the space. Students might count the individual squares (evidence of the single units stage of the *Area Unit Proficiency*), count the number of squares and multiply (evidence of collection of composites stage of *Area Unit Proficiency*) or multiply the given numbers (evidence of mastery of *Formula Proficiency*). As we do not know which strategy was employed on the basis of this data, the responses to these items were modeled as providing evidence for both proficiency variables. The relationship here can be modeled as a disjunctive one. That is, students need proficiency with one or the other of the variables (or both) to be successful. Again, the BN framework allowed for the specification of such a relationship by specifying prior beliefs. Note that other item designs, such as presenting irregular shapes gridded with squares (forcing counting) or only giving dimensions (forcing multiplication) can help these traditional items better discriminate between stages.

Combination of data from multiple sources

Given that rich digital learning environments do not exist in isolation, this project sought to combine evidence from multiple sources. It combined evidence from digital and non-digital activities and traditional test items with new types of evidence. In particular, information such as whether a student uses one tool before another in a game, how often they use a tool, and what formulas they enter in a calculator are not easily combined with information from teacher observations and traditional test items using conventional modeling approaches. The BN modeled here presents a way to make these combinations.

The effect variables helped us model the testlet effect and the shared variance between evidentiary pieces from the same activity. As described above, "effect" is used here as a generic term to describe the association due to context. There are likely a number of issues contained in this variable, including familiarity with the game genre and ease of use of the user interface. Here, we use it to account for the common variance. Future research could continue to disaggregate the variable into its component parts. This effect has been a known issue in traditional testing for decades. Items from one testlet are likely to be more related to each other than they are to items not in the testlet. We would also expect evidence from the same activity to be more related to other evidence from that activity than evidence from another activity. We most often modeled this as a partially compensatory effect, such that context of a task could explain some (but not all) of a lack of proficiency on the proficiency variables. For example, students who are at higher levels of *Area Unit Proficiency* use the Esploda tool more often on the Crab level. However, among those who are novices, the probability of only using Esploda once is .40 for those whose score was lowered in the game context but .27 for those whose score was raised in the game context. The .27 is still higher than the probability for those at higher levels of *Area Unit Proficiency*, but game context does partially compensate for lower levels of proficiency on the math variables. As seen with students 4 and 5 (in Table 4), the effect variables also helped explain the variation in performance across contexts. By modeling the game effect, we could posit the lack of game effect as an explanation for results for student 5, and obtain probabilistic estimates accordingly. Doing so has several benefits. First, latent variables representing the LP are now purified of the influence of the game context. In addition, the results for the context variables may be useful in and of themselves. This information could be shared with teachers as a flag on results. If the probability of being affected reaches a particular threshold, the proficiency result could be flagged as potentially a result of context effects. In this way, students not familiar with game play or tool interactions would be identified and inferences about their math ability could be withheld or tempered.

Synthesizing expert knowledge and empirical data

When creating models of data, it is good to be reminded of Box's (1976) quote, "All models are wrong, some are useful." This work was an attempt to create such a useful probabilistic model that could capture both expert beliefs and data-based relationships. The modeling process included definition of the model such that expert definitions of the stages were reflected. For example, by definition of the stages of the progression, students who have mastered the collection of composites stage should use multiplication on the Hippo activity more often than students who have not mastered composites. In other cases, where hypotheses were not firm, the data either confirmed them or revealed unexpected results. For example, the results for strategy usage in You Kraken Me Up! reported in Table 3 (c) are consistent with the prior belief that the Multi-Gluumi strategy would be more closely associated with being at the single units level of *Area Unit Proficiency* and that the Gluumi-Multi strategy would be more closely associated with being at the collection of composites level of *Area Unit Proficiency*. However, the results also indicate that the Gluumi-Multi strategy is only used by students at the collection of composites level with probability .26. Though this is the largest probability for using that strategy, it was lower than our prior expectations, which were that it would be the most likely strategy used by students at collection of composites level. In contrast to these prior expectations, the results indicate that students at the collection of composites level are more likely to use the Multi-Gluumi strategy. In psychometric terms, this observable is useful for discriminating between students at the single units and collection of composites levels, but not as useful as we had expected it to be. This example highlights how the Bayesian approach allows for the updating of beliefs regarding the conditional probabilities. The posterior for the fitted conditional probabilities of the BN, summarized in Tables 2 and 3, represents our current thinking about the relationships between variables, obtained by synthesizing the information in the data with prior beliefs. In the spirit of Box's dictum, this moves us farther along the road to a useful model for facilitating inferences regarding students (e.g., Table 4 and surrounding discussion). Future work in this area may involve formal statistical procedures to identify strengths and weaknesses of the model (Levy, Crawford, Fay, & Poole, 2011; Sinharay, 2006; Sinharay & Almond, 2007).

The work in this article focused on evaluation of the measurement model, or the links between the evidence and the latent variables. Future research could also use the relationships in the data to explore the structure of the latent variables. For example, the potential correlation between the latent variables could be examined using the patterns of relationships among the evidence pieces. Work of this type could lead to cycles of review and revision of the LPs.

Evaluating models

The validity argument of the model here is strongly dependent on the LP. The BN structure is entirely about modeling the evidence for the LP as it is currently theorized. Our purpose was not to test or judge the theory (or compare models that would tell us about the theory). That could be done, but the first step, shown here, is to build a model that aligns to the LP. This article investigates the ability of the model to provide information about the stages of the LP. The investigation of the LP itself is ongoing, based on both qualitative and quantitative evidence (Lai et al., 2015). That research examines student behaviors and statements to determine whether they indicate the presence of an unmodeled stage or whether there is a lack of evidence for distinguishing between stages. The type of research done here is not well suited for making these conclusions.

One type of validity evidence is comparison to an external measure. Such might be the case if one assessment (e.g., a post-test) might be treated as a target or gold-standard. One might then proceed to build a model excluding that post-test and the compare the model to the scores on that post-test. We have not done that here because we believe that the unique variance of each of the sources is the strength of the model. The goal is not to replicate our existing traditional assessment methods (say, by maximizing prediction of the post-test), but to add more and different evidence to the existing evidence. For that reason, our evaluation of the model above is based on the review of how the results align to our expectations based on the LP.

The conditional probabilities presented here began with the prior beliefs of the experts developing the model. The uncertainty of these beliefs was then incorporated as part of the modeling process. Then, we were able to update the probabilities with data. While larger sample sizes would provide more confidence in our estimates here, the use of any data to update beliefs in the conditional probability table is a step further than many other applications, which tend to model data from a single source relying on conditional probability tables based solely on priors (cf., Iseli, Koenig, Lee, & Wainess, 2010; Rowe & Lester, 2010; Shute & Kim, 2011) or priors and pilot study data (cf., Shute, Wang, Greiff, Zhao, & Moore, 2016). One of the advantages of a BN approach is the ability to continuously iterate and update beliefs as new data become available. Therefore the conditional probabilities presented here are not meant to be the final result, but the result given what we know from the data we have. We believe this should be not just the case when there are 130 participants, but also when there are thousands of participants. Furthermore, as noted above, future work with larger samples may involve formal statistical procedures to identify strengths and weaknesses of the model (Levy et al., 2011; Sinharay, 2006; Sinharay & Almond, 2007).

As it stands, this modeling approach is highly expert dependent for the structure of the BN (which pieces of evidence are related to which levels of the LP). In this case, there was significant research from math educators that led to the LP and suggest the kinds of evidence related to various stages. In newer domains where there is less research or expert opinion is not available, it is likely that the LP should first be developed using methods outside those presented here (for example, cognitive labs). Once the theory of progression is developed, it is possible to use educational data mining techniques to uncover evidence pieces in the activity streams (e.g., Baker, Corbett, Koedinger, & Wagner, 2004; Sao Pedro, Baker, & Gobert, 2012; Wixon, Baker, Gobert, Ocumpaugh, & Bachmann, 2012).

Static vs. dynamic models

This article presents a static model, or a model that is assuming that the learners' proficiency is not changing over time. This assumption is supported by the lack of pre-test to post-test change observed in the students. We can hypothesize a number of reasons for this, including that none of the performance activities were meant to be instructional and that the game did not lead participants up to the formal formula for area (see Lai et al., 2016 for further discussion). This article lays the foundation for the aggregation of diverse evidence from multiple sources when learning is not occurring. The techniques here could certainly be expanded into the framework of dynamic BNs (Reye, 2004; Rowe & Lester, 2010) for situations in which learning is occurring.

Conclusions

The ability to collect activity streams as individuals interact with digital environments provides a record of what students are doing as they progress through learning activities. These data have the potential to yield information about student cognition if methods can be developed to identify and aggregate evidence from diverse data sources. This work demonstrates how data from multiple carefully designed activities can be used to support inferences about students' levels of understanding.

References

Almond, R. G., Mislevy, R. J., Steinberg, L., Yan, D., & Williamson, D. (2015). *Bayesian networks in educational assessment*. New York, NY: Springer.

Almond, R. G., Mulder, J., Hemat, L. A., & Yan, D. (2009). Bayesian network models for local dependence among observable outcome variables. *Journal of Educational and Behavioral Statistics*, *34*, 491–521. doi:10.3102/1076998609332751

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behavior in the cognitive tutor classroom: When students "Game The System". In *Proceedings of ACM CHI 2004: Computer-Human Interaction*, pp. 383–390. doi:10.1145/985692.985741

Barrett, J. E., Cullen, C., Sarama, J., Clements, D. H., Klanderman, D., Miller, A. L., & Rumsey, C. (2011). Children's unit concepts in measurement: A teaching experiment spanning grades 2 through 5. *Zdm*, *43*(5), 637–650. doi:10.1007/s11858-011-0368-8

Battista, M. T., Clements, D. H., Arnoff, J., Battista, K., & Borrow, C. V. A. (1998). Students' spatial structuring of 2D arrays of squares. *Journal for Research in Mathematics Education*, *43*, 503–532. doi:10.2307/749731

Baturo, A., & Nason, R. (1996). Student teachers' subject matter knowledge within the domain of area measurement. *Educational Studies in Mathematics*, *31*(3), 235–268. doi:10.1007/BF00376322

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. doi:10.1080/01621459.1976.10480949

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168. doi:10.1007/BF02294533

Collins, A., & Halverson, R. (2009). *Rethinking education in the age of technology: The digital revolution and schooling in America*. New York, NY: Teachers College Press.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform (No. CPRE research report #RR-63)*. Philadelphia, PA: Consortium for Policy Research in Education.

DiCerbo, K. E., & Behrens, J. T. (2012). Implications of the digital ocean on current and future assessment. In R. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 273–306). Charlotte, North Carolina: Information Age Publishing.

DiCerbo, K. E., & Behrens, J. T. (2014). *The impact of the digital ocean on education. [white paper]*. London, England: Pearson. Retrieved from https://research.pearson.com/digitalocean

DiCerbo, K. E., Crowell, C., & John, M. (2015). Alice in Arealand. Worked example presented at games+learning +society conference, Madison, WI. In K. E. H. Caldwell, S. Seyler, A. Ochsner, & C. Steinkuehler (Eds.), *GLS11 conference proceedings* (pp. 382–385). Madison, WI: Games+Learning+Society. Retrieved from http://press.etc.cmu.edu/content/gls-11-conference-proceedings

Iseli, M. R., Koenig, A. D., Lee, J. J., & Wainess, R. (2010). *Automatic assessment of complex task performance in games and simulations* (CRESST Research Report No. 775). Los Angeles: National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved from http://www.cse.ucla.edu/products/reports/R775.pdf

Keller, L. A., Clauser, B. E., & Swanson, D. B. (2010). Using multivariate generalizability theory to assess the effect of content stratification on the reliability of a performance assessment. *Advances in Health Sciences Education*, *15*(5), 717–733. doi:10.1007/s10459-010-9233-8

Lai, E., Kobrin, J., & DiCerbo, K. E. (2016). *Developing and piloting the insight learning system*. Paper presented at the annual meeting of the National Council for Measurement in Education, Washington, D.C.

Lai, E. R., Kobrin, J., Nichols, P., & Holland, L. (2015, April). *Developing and evaluating learning progression-based assessments in mathematics*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Levy, R., Crawford, A. V., Fay, D., & Poole, K. L. (2011, April). *Data-model fit assessment for Bayesian networks for simulation-based assessments*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Levy, R., & Mislevy, R. J. (2004). Specifying and refining a measurement model for a computer-based interactive assessment. *International Journal of Testing*, *4*(4), 333–369. doi:10.1207/s15327574ijt0404_3

Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling*. Boca Raton, FL: Chapman & Hall/CRC.

Mayer-Schonberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think* (1st ed.). Boston, MA: Eamon Dolan/Houghton Mifflin Harcourt.

Mislevy, R. J., Oranje, A., Bauer, M. I., vonDavier, A., Hao, J., Corrigan, S., … John, M. (2014). *Psychometric considerations in game-based assessment. [white paper]*. Retrieved from Institute of Play website http://www.instituteofplay.org/work/projects/glasslab-research/

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3–62.

NPD Group. (2011). *The video game industry is adding 2–17-year-old gamers at a rate higher than that age group's population growth*. New York, NY: Author.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*, 667–696. doi:10.1080/00273171.2012.715555

Reye, J. (2004). Student modelling based on belief networks. *International Journal of Artificial Intelligence in Education*, *14*, 63–96.

Rowe, J. P., & Lester, J. C. (2010). Modeling user knowledge with dynamic Bayesian networks in interactive narrative environments. In G. M. Youngblood & V. Bulitko (Eds.), *Proceedings of the sixth AAAI conference on artificial intelligence and interactive digital entertainment, AIIDE 2010*. Retrieved from http://aaai.org/ocs/index.php/AIIDE/AIIDE10/paper/view/2149

Rupp, A. A., & Leighton, J. P. (Eds.). (2016). *Handbook of cognition and assessment*. Malden, MA: John Wiley & Sons.

Sao Pedro, M. A., Baker, R. S., & Gobert, J. D. (2012). Improving construct validity yields better models of systematic inquiry, even with less information. In J. Masthoff, B. Mobasher, M. C. Desmarais, & R. Nkambou (Eds.), *Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization* (pp. 249–260). Berlin, Germany: Springer.

Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research: Learning trajectories for young children*. New York, NY: Routledge.

Shute, V. J., & Kim, Y. J. (2011). Does playing the world of goo facilitate learning. In D. Y. Yai (Ed.), *Design research on learning and thinking in educational settings: Enhancing intellectual growth and functioning* (pp. 359–387). New York, NY: Routledge.

Shute, V. J., Leighton, J. P., Jang, E. E., & Chu, M.-W. (2016). Advances in the science of assessment. *Educational Assessment*. doi:10.1080/10627197.2015.1127752

Shute, V. J., Wang, L., Greiff, S., Zhao, W., & Moore, G. (2016). Measuring problem solving skills via stealth assessment in an engaging video game. *Computers in Human Behavior*, *63*, 106–117. doi:10.1016/j.chb.2016.05.047

Sikorski, T. R., & Hammer, D. (2010, June). A critique of how learning progressions research conceptualizes sophistication and progress. *In Proceedings of the 9th International Conference of the Learning Sciences-Volume 1*, (pp. 1032–1039). International Society of the Learning Sciences.

Simon, M. A., & Blume, G. W. (1994). Building and understanding multiplicative relationships: A study of prospective elementary teachers. *Journal for Research in Mathematics Education*, 472–494. doi:10.2307/749486

Sinharay, S. (2006). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics*, *31*, 1–33. doi:10.3102/10769986031001001

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models: A case study. *Educational and Psychological Measurement*, *67*, 239–257. doi:10.1177/0013164406292025

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*, 237–247. doi:10.1111/jedm.1991.28.issue-3

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2007). *WinBUGS [computer software]*. Cambridge, UK: MRC Biostatistics Unit.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.

West, P., Rutstein, D. W., Mislevy, R. J., Liu, J., Levy, R., DiCerbo, K. E., … Behrens, J. T. (2010). *A Bayesian network approach to modeling learning progressions. CRESST Research Report*. Los Angeles, CA: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved from http://www.cse.ucla.edu/products/download_report.asp?r=77

Wixon, M., Baker, R. S. J. D., Gobert, J., Ocumpaugh, J., & Bachmann, M. (2012). WTF? Detecting students who are conducting inquiry without thinking fastidiously. In J. Masthoff, B. Mobasher, M. C. Desmarais, & R. Nkambou (Eds.), *Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization (UMAP 2012)* (pp. 286–298). Berlin, Germany: Springer.