Measurement, 12: 1–33, 2014

ISSN: 1536-6367 print / 1536-6359 online
DOI: 10.1080/15366367.2014.910060

Routledge
Taylor & Francis Group

FOCUS ARTICLE

How Task Features Impact Evidence From Assessments Embedded in Simulations and Games

Russell G. Almond, Yoon Jeon Kim, Gertrudes Velasquez, and Valerie J. Shute
Educational Psychology and Learning Systems, Florida State University

One of the key ideas of evidence-centered assessment design (ECD) is that task features can be deliberately manipulated to change the psychometric properties of items. ECD identifies a number of roles that task-feature variables can play, including determining the focus of evidence, guiding form creation, determining item difficulty and discrimination, characterizing proficiency, and producing task variants. Assessment developers can use these task features to manipulate the psychometric properties of both conventional assessment formats and complex tasks embedded in simulations and games.

Simulations and games tasks present additional challenges: even defining what corresponds to an item can be difficult. Often task features are determined by game or simulation logic rather than psychometric design. Despite these difficulties, the roles for task features identified in ECD are useful for analyzing the psychometric properties of embedded assessments in simulations and games. This article compares a conventional test of mathematics-problem-solving ability using word problems to an assessment of conceptual physics, creativity, and conscientiousness embedded in the game *Newton's Playground,* describing how the task roles play out in each setting and how they can be used to manipulate the evidence provided by an assessment.

Keywords: evidence-centered assessment design (ECD), games and assessment, task models, difficulty, discrimination, test assembly

EXTENDING EVIDENCE-CENTERED ASSESSMENT DESIGN TO EMBEDDED ASSESSMENTS

Educators are increasingly turning to simulations and games to assess knowledge, skills, and abilities that are difficult to address with more-familiar item types (particularly, multiple

Correspondence should be addressed to Russell G. Almond, Educational Psychology and Learning Systems, Florida State University, 3204-J Stone Building, 1114 W. Call St., Tallahassee, FL 32306-4453. E-mail: ralmond@fsu.edu

choice). However, the language of assessment design has grown up around multiple-choice and short-answer items, and it is often not easy to express the design of tasks embedded in more-complex activities.

Shute, Ventura, Bauer, and Zapata-Rivera (2009) call these kinds of assessments embedded in games and simulations *stealth assessments,* because they blur the line between assessment and learning activity. Players focus on the activity and not the fact that their abilities are being assessed. Note that a game-based assessment may or may not be an *embedded assessment* in the sense of Dori (2003)—that is, a formative assessment that is integrated within a curriculum to monitor students' progress. Although alignment between assessment task and curriculum is critically important, the focus of this article is on providing a language to help assessment designers with designing tasks for specific purposes and thinking about which portions of the task can be manipulated in the design process and to what purpose.

Evidence-centered assessment design (ECD; Mislevy, Steinberg, & Almond, 2002, 2003) was developed in part to provide assessment designers with ways of talking about the new kinds of concerns raised by new kinds of assessments. To wean assessment designers away from assumptions that are not necessarily true in new types of assessments, ECD deliberately introduced new language. One shift was replacing the term "item" with "task," the new term is meant to remind designers that activities embedded in games and simulations are often extended and yield complex work products that generate multiple observed outcomes. For example, a task might consist of reading a passage of text and then answering several questions (traditional items, each producing a different observed outcome), or the task might consist of carrying out several activities in a game (here the observed outcomes are based on the game activity sequence and end state).

The second shift was to move from talking about individual tasks to talking about *task models*—a universe of possible tasks that share similar features and provide similar evidence regarding claims about the examinees to be determined by the assessment. In particular, a task author could deliberately manipulate features of a task—*task model variables*—to change the evidential features of the task. Mislevy et al. (2002) define a number of different ways that assessment designers can use the task model variables.

Depending on the format of assessment, however, how designers consider roles of task model variables will vastly differ. To that end, the article will demonstrate the use of task model variables in two settings. The first is the mathematics word problem, a problem type that is used throughout elementary mathematics education. The second is a game, *Newton's Playground,* which is designed to contain an embedded assessment of the player's knowledge of conceptual Newtonian mechanics, as well as embedded assessments of creativity and conscientiousness (Shute & Ventura, 2013). The next section describes the two applications and is followed by a brief review of ECD. The remaining sections explore the various roles for task model variables identified in Mislevy et al. (2002) and provide examples in both contexts.

TWO RUNNING EXAMPLES

To illustrate the manipulation of task model variables, this article focuses on two problem types. The first is the mathematics word problem, which is frequently used in K–12 education. In this

problem type, the prompt is a short (about one-paragraph) story that contains an embedded problem. The expected response is a number of mathematical expressions. This kind of task can be presented either on a computer (with or without automatic scoring) or on pencil and paper. The second example is the game *Newton's Playground*. This game was designed with the twin goals of both being enjoyable to play and containing a *stealth assessment*—a valid assessment embedded seamlessly within the game environment (Shute et al., 2009). ECD is the framework used to link the observed behaviors in the game to the claims about proficiency the designers wish to make about the player. One facet of this is to deliberately manipulate task model variables to make the challenges within the game provide better evidence of the targeted competencies. The use of two task types, one familiar and one unfamiliar, is intended to make the distinctions between conventional and game-based assessments clearer.

Mathematics Word Problems

Word problem solving provides students with intellectual challenges to enhance their mathematical computation while fostering reasoning and mathematical communication abilities (Hiebert & Wearne, 1993). The National Council of Teachers of Mathematics (NCTM, 1989) considers teaching problem-solving skills as a means of teaching mathematics by providing students with opportunities to "acquire ways of thinking, habits of persistence and curiosity, and confidence in unfamiliar situations" (p. 52). Solving arithmetic and algebra word problems involves associating the context in which information is presented with a recognized, albeit idealistic, real-world scenario and retrieving embedded information to form mathematical relationships among the data.

Recognition of contextual information results in classification of problems into different types. For example, arithmetic story problems can be categorized into 5 basic problem structures (*change, combine, compare, vary and transform*) based on schema theory (Marshall, Pribe, & Smith, 1987), or combinations of these. Similarly, introductory algebra word problems can be classified as number problems, business/investment problems, uniform-motion problems, and mixture problems. Recognition of these categories assists in formulating relationships among the data and developing methods for obtaining solutions. Ability to solve these basic types of word problems becomes the foundation for development of abstract-thinking skills (Gick & Holyoak, 1983; Morales, Shute, & Pellegrino, 1985; Cooper & Sweller, 1987).

Those characteristics of word problems can guide task development in educational assessments, specifically, in the development of *task shells* for categories of word problems. The stem of a task shell consists of a story with a number of blanks. These blanks are variables that are filled in either manually by the task author or automatically from a set of approved choices. The key is a function of certain combinations of the variables. The work product of the task is either a short answer or a selection from a list containing the correct answer and a number of distractors (multiple choice). In the latter case, distractors are also often chosen as functions of the variables (Bart, Post, Behr, & Lesh, 1994; Graf, 2008).

As an illustration, the following task shell involving uniform motion will be used to illustrate the different roles of task model variables:

_____ _____ _____ _____ at _____
(Name of actor) (verb of motion) (a (un)known distance) (unit of distance) (a (un)known speed)
_____ _____ _____ _____ _____
(unit of speed). (Then/thereafter) (pronoun for actor) (different verb of motion) (a second (un)known distance)
_____ at _____ _____. The _____
(second unit of distance) (a second (un)known speed) (second unit of speed) (total/difference between)
_____ was _____ _____. What is the _____ the _____
(time) (a quantity for time) (unit of time) (distance traveled in) (first part/second part/total)
of the journey?

This task shell for uniform motion can be used to develop specific tasks at different levels of complexity. Specific examples are provided below:

Lower middle school arithmetic story problem: Baby Jasmine walked 10 paces. She then ran 13 paces. What is the total number of paces that Baby Jasmine Traveled?

Middle school algebra problem: On a certain journey, Aaron walked x miles at 2 miles per hour. Then, he cycled 9 miles at 3 miles per hour. The total time for the journey was 5 hours. What is the distance traveled in the first part of the journey?

Introductory algebra problem: On a certain journey, Annmarie ran x km at 2km/h. She then cycled $3x$ km at 4km/h. The total time taken for the journey was 7 hours. What is the distance traveled in the first part of the journey?

A more complex algebra problem: Eric drove a certain distance at 40 mph in the first part of his journey. He then drove the same distance at 20 mph faster than the former speed. The total time taken was 0.5 hours. What is the distance traveled in the first part of the journey?

To author a task from this shell, the blanks are filled in according to a set of rules (with minor wording changes to maintain grammatical correctness). Note that often there are constraints among the possible values. For example, if the verb is "walking" then "km/s" would not be an appropriate unit of speed. A second constraint is that the units of distance and speed are the same in both contexts (otherwise an additional unit conversion subtask is necessary). Other constraints on the variables are based on the evidentiary purpose of the tasks. For example, if the initial distance is represented with a symbol instead of a number, that changes the problem from an arithmetic problem to an algebra problem. Whether or not a given constraint is satisfied as a task model variable, these kinds of constraint variables often play a role in determining the psychometric properties of the tasks as described below.

The task shell presented here is a special case of the more general *task template* (Riconscente, Mislevy, & Hamel, 2005). The task template goes beyond the simple shell, providing the task author with guidance about how to use the task variables to produce tasks for particular purposes.

Newton's Playground

Newton's Playground is a 2-dimensional physics game. The mechanics of the game are fairly simple. The game is divided into a series of levels. In each level the player is presented with a line drawing containing both fixed and movable objects. The drawing always contains a ball (the focus of play) and a balloon (the goal). The player draws objects such as ramps, levers, pendulums, and springboards or just adds additional masses exerting a force to the ball to move the ball to the balloon(s). All drawn and preexisting objects (which are not fixed in place) follow Newton's laws of motion, with gravity exerting a force directed toward the bottom of the screen

on all nonfixed objects. The physics simulation is done using the open source Box 2D (Catto, 2011) physics engine for games.

Figure 1 shows the initial configuration of a level called Spider's Web. The ball is below the balloon target, so the player needs to add potential energy to the ball to reach the balloon. Figure 2 shows one potential solution. The player has added a springboard to the ledge using 2 pins (small round circles). The player has also attached a large weight to the springboard, pulling it down. Deleting the weight will cause the ball to fly up in the air. The ramp attached to the top of the figure will keep the ball from flying over the target.

The player can spend as long as necessary to complete a level. If the ball falls off the screen, it is replaced with a new one. Objects can be deleted and redrawn. If the player gets stuck, he or she can restart the level from the beginning. The player can also solve the level many times, earning

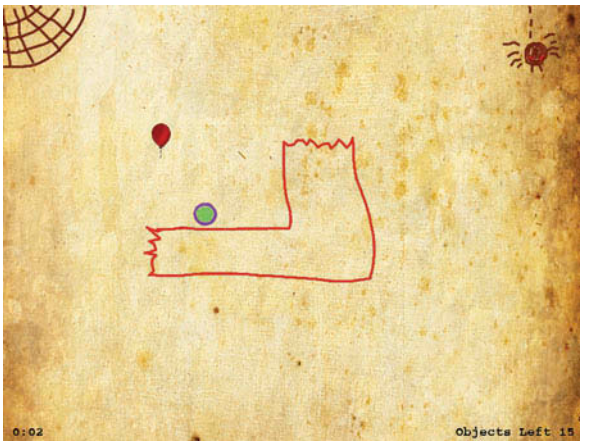

FIGURE 1 Starting position for Spider Web level (color figure available online).

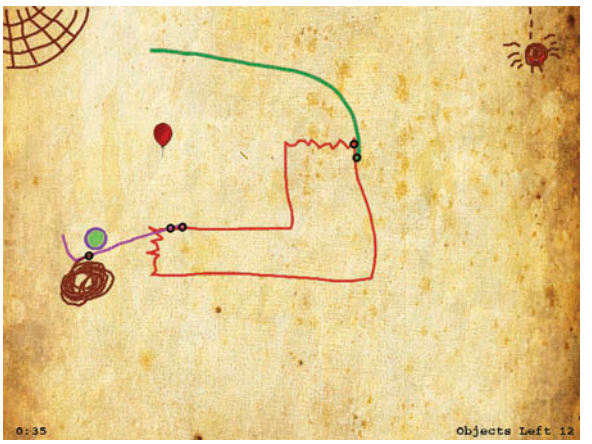

FIGURE 2 Spider Web level with springboard solution (color figure available online).

an additional trophy each time the level is solved. The levels come in sets called playgrounds, and the player can skip around between levels, abandoning a difficult level and returning to it later.

For most levels, there are many possible solutions, although some may be more obvious than others. The game development has focused on levels that can be solved using 4 agents of motion: ramps, levers, pendulums, and springboards. Some levels are designed to focus on one agent, while others are more open ended.

Even though *Newton's Playground* is a casual game that people of all ages can play for entertainment, it was specifically developed to assess conceptual physics understanding, creativity, and conscientiousness (particularly persistence). When students play the game, the game records a trace of all their activities including all the objects that the students drew and saves the data in the log files, which can be replayed using a built-in replayer. The system also records information about how much time was spent on each level and how many successful solutions the player made during that time.

A BRIEF REVIEW OF EVIDENCE-CENTERED ASSESSMENT DESIGN

The discussion of task design in the following sections will make use of the language of evidence-centered assessment design. This section provides a brief review, focusing on those parts of ECD relevant to the discussion of roles for task model variables, and is not a complete description. For readers unfamiliar with ECD, the article by Mislevy, Steinberg, Almond, and Lukas (2006) provides a good starting point.

ECD, as described in Mislevy et al. (2003), is simultaneously a process for designing assessments, a set of principles for addressing the design decisions that arise, and a formal language for describing the design. The latter is called the *conceptual assessment framework* (CAF, Figure 3),

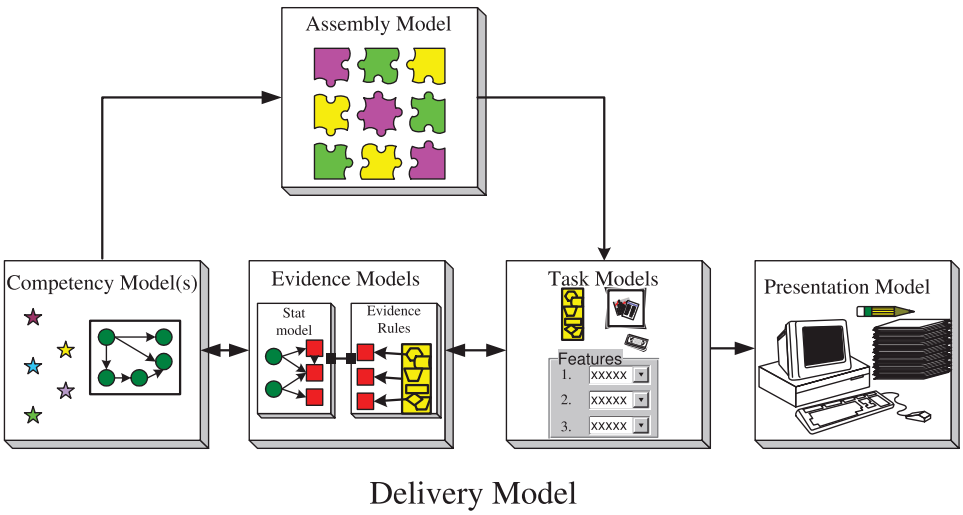

FIGURE 3 The principle design objects of the CAF
Source: Mislevy et al., 2006. (color figure available online).

and this section will briefly review the principle design objects of the CAF, called *models*. One way to define the ECD models revolves around a series of questions that must be answered (explicitly or implicitly) in any assessment design. The questions are identified and addressed, in turn, in the following sections.

What collection of claims about students is being measured? (Student) competency models

The stars in the box in Figure 3 represent claims the designers would like to make about examinees. Often these claims are organized into one or more *competency variables*; the reported scores are usually statistics of these variables (or their distribution). The competency model describes the distribution of values for these competency variables in the population. When scoring an individual, the scoring generally makes a student-specific copy of the competency model (a *student model*), which keeps track of what is known about that individual's performance.

If the ECD measurement model is Bayesian, the competency model is the prior distribution over the possible configurations of the competency variables, based on the target population of the assessment. The student model is the posterior distribution for a particular student, the distribution after absorbing the evidence from the observed outcomes of the tasks attempted by the student.

Note that ECD explicitly assumes that competency may be multidimensional. There are a large number of additional design issues that arise when a task can potentially address one or more competency variables. These issues are often irrelevant when the competency model is unidimensional, and assumptions about unidimensionality are often compiled into assessment-design procedures. The language of ECD helps designers work through issues arising from multidimensional-competency models, which often add confusion to the design.

For each of the examples in the previous section, there are a number of possible choices for the competency model; for the word problem example, the primary claims of interest relate to the ability to solve word problems and the specific requirements for these claims will vary depending on grade level. For example, the Common Core State Standards for 1st-grade mathematics include, "Solve word problems that call for addition of three whole numbers whose sum is less than or equal to 20" (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010, CCSS.Math.Content.l.OA.A.1), while the 4th-grade standards include, "Solve multistep word problems posed with whole numbers and having whole-number answers using the four operation" (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010, CCSS.Math.Content.4.OA.A.3).

We will place those claims on a continuous scale we will call θ_M. To make the scale of the variable identifiable, we will assume that the values of θ_M are normally distributed with a mean of 0 and a standard deviation of 1 in the target population. A complication arises from a lurking secondary competency in the word problem example, the (language) ability to read and comprehend the story. In particular, the linguistic complexity of math word problems influences item difficulties (Martiniello, 2008). There are several possible ways to handle this, including (1) assuming that language ability is part of the measured competency θ_M and creating a unidimensional-competency model and (2) introducing a new variable, θ_L, to represent language competency. As the latter approach offers more flexibility, that is the one followed in this article. With this choice, θ_L is assumed to have a normal distribution in the target population (mean, zero; standard

deviation, one). Additionally, we must also specify a correlation between θ_M and θ_L, possibly based on the observed value in another study or in a pretest population.

The *Newton's Playground* competency model is complex. The designers wish to make claims about understanding of physics, conscientiousness (particularly, persistence), and creativity. Each of those constructs is itself complex, and described with multiple variables. A discrete Bayesian network is used to describe the joint distribution of the variables, with various claims attached to various levels of the variables. Again there are design choices to be made. The 3 domains of the assessment could be represented with 3 different competency models (with 3 separate Bayesian networks) or a single big competency model that encompasses all 3 domains. As there are likely interactions of the competencies in the problems (e. g., a greater knowledge of physics will make the player more fluent in generating creative solutions), the latter approach is the one that was taken for *Newton's Playground.*

What kinds of observations provide evidence about the measured claims? Evidence models

In ECD, evidence is processed in 2 phases. First, the *rules of evidence* process the raw *work product* of the task (the complicated drawing on the right side of the evidence model) and calculate the values of certain key *observable outcome variables.* Next, the *statistical part* of the evidence model describes the relationship between the competency variables and the observable variables and how a student's competency model should be updated to reflect the new evidence. (In Bayesian ECD, the statistical part is the likelihood of the observations given the competency variables, and the update rule is Bayes's theorem applied using the student's current competency distribution as a prior.)

Consider the word problem example. The rules of evidence will depend on whether the task outcomes are to be scored by human or machine. For human scoring a set of rules, *a rubric,* describing what constitutes a correct answer must be created (possibly with blanks to be filled in based on the values of the task model variables). The rules of evidence must address issues such as whether units are required, how accurate the answer needs to be, and whether algebraic expressions need to be simplified. If machine scoring is to be used, then the rules of evidence need to be realized as computer code. In either case, the observable variable, Y_{ij}, is set to 1 if the student entered a string that meets the criteria for a correct answer and to 0 if the student entered a string that does not. There are a number of possible choices for the statistical part of the evidence model. A commonly used choice is a compensatory 2-parameter logistic item response theory (IRT) model, $\Pr Y_{ij} = 1 = \log \mathrm{it}^{-1}\left(a_{Mj}\theta_{Mi} + a_{Lj}\theta_{Li} - b_j\right)$, where b_j is a task-specific difficulty parameter and a_{Mj} and a_{Lj} are task-specific discrimination parameters.

In *Newton's Playground* both the rules of evidence and the statistical part of the evidence model are more complex. One observable feature of a solution is whether or not the player used a lever as part of the solution. The rules of evidence need to be able to identify what kind of drawing on the screen constitutes a lever (an object that rotates but does not have a fixed pivot that would make it a pendulum). This could either be expressed as a rubric for a human rater evaluating a solution (especially during early phases of development) or computer software that processes the replay information. The game also distinguishes between an efficient solution (one with few drawn objects), for which a gold trophy is awarded; a less efficient solution, for which a

silver trophy is awarded; and no solution. As a result, one of 4 observable variables corresponding to the 4 agents of motion is set to a value that indicates what kind of trophy was earned, while the others are said to be unobserved. In the case of no solution, only the variables thought to be reasonable solution paths by the designers are set. As sufficient information to completely replay the level is stored by the game, other observable variables are possible, but their use has not yet been explored.

The observation that a player solved a level using a lever provides evidence in 2 different ways. First, because operating the lever involves both linear and angular momentum, the use of a lever provides evidence that the student has some understanding of those concepts (and the probability of lever use is higher for students with higher levels of conceptual physics understanding). Second, if the game level is one that is most often solved using a springboard, the use of a lever is an indication of creativity.

The statistical part of the evidence model for *Newton's Playground* is a Bayesian network fragment linking the observable outcome variables for a specific level to the competency model variables. Constructing this network fragment requires 2 steps. First, the graphical structure of the model must be determined. In particular, the game designers must specify which competency variables are relevant for which observables and add the corresponding edges to the network. Second, conditional probability tables need to be established for each observable variable. Almond et al. (2001) suggest taking standard models from IRT and discretizing them. Note that this approach uses difficulty and discrimination parameters similar to those used in the IRT model above. Almond, Kim, Shute, and Ventura (2013) describe the process of constructing the statistical model for *Newton's Playground* in more detail.

How do we engineer settings to provide the required observations? Task models

From an ECD perspective, a task is some kind of activity that is attempted by the examinee and yields a distinct *work product* (which can be processed by evidence rules). Usually there is some kind of material that is present to the examinee; e. g., instructions, prompt stem, key and distractors (for multiple choice tasks), pictures, audio, and video.

Note that in the context of games and simulations it can be difficult to set the boundaries of what a task is. In *Newton's Playground,* game play is divided into a number of levels, each of which represents a distinct puzzle. The player can play the level until solved or abandon it and move to another level (possibly returning later). The player can also solve a single level multiple times, each with a different solution. One question that arises is what should we consider to be a task (in the ECD sense). A single attempt? Multiple attempts until the player moves to another level? All of the attempts within a single session of game play? These design decisions have impact on what can be considered an observable, and when the system has sufficient evidence to provide scores or other forms of feedback based on competency estimates.

Usually, all of the observables arising from the work product of a single task are scored using a single evidence model (or set of evidence models). As ECD is based on a relaxed version of the local independence assumption of many IRT models, observables coming from different tasks are assumed to be independent given the competency variables; however, the design can explicitly model the dependencies coming from observables from the same task. On the basis of this rule, the best definition of "task" for *Newton's Playground* appears to be all attempts at the same level

within a single session of game play. Other games and simulations may require more complex definitions, especially if the activities are not so easily broken down into levels or if the system needs to adapt to player ability within the game or simulation.

A task model is an abstract description of a collection of related tasks. It describes what kind of presentation material is used and what kind of work product is expected (in what format). It also puts constraints on the allowable presentation material. Task models also define *task-model variables* appropriate to the class of tasks they support. Task model variables are features of the presentation material or expectations about the work product, and their value in a particular task will be tied to choices made about the presentation material.

For example, consider a task model variable that identifies the vocabulary level of several paragraphs of text that need to be read as part of the task. This could be used in several ways: (1) The task author could write the text and then compute the vocabulary level. If the authored material does not meet the bounds set by the task model, the material would be edited; (2) the task author could search a corpus of text for a passage that met the vocabulary requirements (and other constraints defined in the task model); (3) the text could be generated automatically, with words chosen from a word list at the appropriate vocabulary level.

Identifying task model variables helps authors identify parts of tasks that can be manipulated for various effects. For example, consider the task model variable from *Newton's Playground* that records whether the balloon (goal) is higher or lower than the initial position of the ball. If the balloon is higher, this changes the nature of the problem, as the player must provide additional potential energy to the system for the ball to reach the goal. Once the designers became aware of this variable, they could deliberately manipulate it to achieve a variety of goals (This is discussed in the section entitled "Roles for Task Model Variables"). Task templates combine this kind of guidance with the task model.

How much evidence is required to adequately support/refute the claims? Assembly model

While the competency model lays out the intended meaning of the competency variables in terms of the claims, the collection of tasks used to measure that competency defines its effective meaning. Consider a proficiency variable, *understands graphs and tables.* If all of the tasks used graphs and none used tables, then the effective meaning of the variable would not match its label. Thus, the mix of tasks that goes into the assessment is an important aspect of its validity.

ECD's approach to specifying valid assemblies is best understood in terms of the problem of assembling an optimal test form from a bank of items using combinatorial optimization (van der Linden, 2005). This approach starts by identifying a target or objective function that essentially states how much evidence is needed (for each variable in the multidimensional case). This is balanced by a series of constraints on the test form such as length, expected completion time, the tasks' spanning the construct being assessed, and the tasks' being appropriately distributed among forms. (i.e., not violate the local independence assumptions).

For multidimensional assessment, the Q-matrix (Rupp, Templin, & Hensen, 2010; Tatsuoka, 1983) emerges from the interaction of the assembly and evidence models. The assembly model defines which tasks appear on the assessment. Those tasks (or the observables associated with the tasks) are the rows of the Q-matrix. The competency variables referenced by the evidence

models are the columns of the Q-matrix. In general, the value $q_{jk} = 1$ when Observable j depends on Competency k, and $q_{jk} = 0$ if it does not. From this pattern of 1s and 0s, various properties of the instrument can be derived (Leighton & Gierl, 2007). The Q-matrix also provides a useful tool for assessment designers, and often the evidence and assembly models can be generated from a Q-matrix, augmented with additional information about the tasks (Almond, 2010).

In the case of the math word problem test, the objective function relates to the desired test-information function (Hambleton, Swaminathan, & Rogers, 1991). Trying to maximize this constraint leads to a spread of difficulties. But other constraints are needed as well, to ensure that the full spread of contexts is explored (i.e., spanning the construct), also that the variants are not so close that one might provide a clue to the solution of the other. This latter would be a threat to the local independence assumption required by many psychometric models.

The assembly model presents some difficulty in assessment embedded within games and simulations. First, the desire to gain specific kinds of evidence needs to be balanced with the internal logic of the game play—it cannot get in the way of the player having fun! In a school setting, the student can simply be told to answer all of the problems on a test.

Often, the player is given much more autonomy in a game. That is, in a game or simulation, players can repeat tasks many times or avoid some tasks that would provide evidence that the designers want. There may be creative solutions to these problems, such as offering greater in-game rewards for tasks that would provide more information, but many of these have yet to be tried.

In *Newton's Playground,* the game levels were grouped into "playgrounds," with the idea that players would attempt the tasks within one playground within one session of play. Thus, the designers can arrange tasks within the playground to achieve certain evidentiary targets much the way that traditional test designers arrange items on a form. However, as players can skip around within the playground, and can spend as much or as little time per level as they like, it is unknown whether this will provide a good balance of information. Play testing here is important to make sure these goals are achieved—and to strike the right balance between learning and fun.

How are the tasks rendered in different physical (or electronic) environments? Presentation models

Think about the way that a task might be rendered in a Web browser. The task is like the HTML file which the browser initially received; it provides the outline of the task and contains links to other presentation material and opportunities for inputs. The presentation model is like a style sheet for that Web page: it shows how the presentation material and input controls should be rendered on different platforms. Test designers have always found this separation of content specification and display instructions useful for tests that must be delivered in both paper-and-pencil and computer environments and will find it increasingly useful as testing expands to new platforms such as tablet computing and smart phones. It also allows test developers to extend the concept of universal design to assessments and to let them reason about how to best accommodate examinees with special needs (Hansen & Mislevy, 2004).

With the math word test example to be administered via the Internet, the presentation model should include specifications regarding interface design, colors, fonts, and so on, as well as what controls are available for entering the solution. Now consider an alternative presentation

of the task on a mobile device. A different presentation model would be required to specify new choices of font, layout, and interaction capabilities.

In simulation- and game-based assessment, the presentation model also includes the rules of the game and the mechanics of the simulation. In fact, earlier ECD papers (e.g., Mislevy et al., 2002) referred to the presentation model as a simulation model. This is important because there needs to be a compelling need for the player to produce the evidence needed for the assessment that does not interrupt the flow of the game. For example, a game in which a player has to stop in the middle of a fire fight to do math problems to recharge the player's blaster breaks the suspension-of-disbelief for the game, and students forced to play the game will judge it to be "lame."

The simulation engine also provides important constraints on the kinds of tasks that are available. *Newton's Playground* is based on the Box 2D physics engine (Catto, 2011). This provides both opportunities and constraints. For example, all of the tasks in *Newton's Playground* have set gravity to Earth-normal conditions. However, if desired, new levels could be made using lunar or even microgravity conditions; of course, these new levels would address slightly different competencies from the current game. A limitation of the current *Newton's Playground* implementation is that there is no visible indication of physics concepts (e.g., forces, velocities, momentum) on the screen other than the behavior of the primary objects over time. This makes it very difficult to gather evidence about how much a player knows about the formal language of physics, as opposed to using physics principles and simple machines to solve mechanical problems.

What other design constraints bear on the assessment? Delivery system model

There are a large number of important design issues that belong to the assessment as a whole, and not to one of the smaller models; for example: How long will the assessment take? What concerns are there for security? Will skipping of tasks be allowed, and how will skipped tasks be scored? All of these serve as constraints on the design that must be respected by the other models.

A key principle of simulation and game-based assessment is that the assessment environment must be engaging (Shute et al., 2009). The goal is not to hide the fact that the players are being assessed but rather to have them lose their self-consciousness about being assessed and devote themselves fully to solving the problems in the game. Under this situation the assessment should yield the maximum evidence (Wise, Wise, & Bhola, 2006).

Gee (2010) noted that successful games do a good job of controlling the difficulty of challenges. A good game proceeds in a series of challenges, each one more difficult than the previous one and each one teaching the player something more about the mechanism of the game. This is very similar to optimizing an assessment to provide maximum information and is also a condition under which maximum learning occurs (Shute, Hansen, & Almond, 2008). Therefore, by learning to manipulate the psychometric properties of challenges, designers can optimize a game for measurement, learning, and fun.

ROLES FOR TASK MODEL VARIABLES

Evidence-centered assessment design takes its name because its notion of assessment design is based on a fundamental evidentiary argument: *We (the testing authority) believe that these claims*

do (not) hold for this examinee because we observed this kind of evidence in this set of contexts. The claims come from the competency model. The observations come from the evidence model, the specific contexts come from the task model, and the set of contexts comes from the assembly model. The goal of the assessment designer is to create a set of contexts for making observations that support the inferences of the assessment.

Mislevy et al. (2002) laid out a number of mechanisms that could be used to manipulate the psychometric properties of tasks. The idea is that task model variables play 1 or more roles in the design of the task, and deliberately manipulating them could produce tasks that meet various evidentiary needs. (Similarly, examining their value for naturally occurring tasks, such as levels in an existing game, could help designers understand what evidence could be gleaned from that task.) The following roles are the ones identified by Mislevy et al. (2002), slightly reorganized:[1]

1. *Creating Task Variants.* Manipulating variables that have this role will make the task appear different to the examinee.
2. *Focusing Evidence: Driving Difficulty.* Manipulating variables that have this role will change the level of competency required to have a high probability of making a particular observation. If IRT models are built for the observable, task variables with this role will shift the location of the curve.
3. *Focusing Evidence: Evidential Strength (Discrimination).* Manipulating variables that have this role will make the correlation between the targeted competency and the observation stronger or weaker. If IRT models are built for the observable, task variables that have this role affect the slope of the curve (discrimination).
4. *Focusing Evidence: Relevant Competencies.* Variables that have this role control which competency variables are relevant for a particular observation. If a *Q*-matrix is built for the assessment, these variables control the pattern of 0s and 1s. Equivalently, if a structural equation model or a Bayesian network is used, these variables control which competency variables are connected to which observables.
5. *Characterizing Proficiency.* In ECD, proficiency or competency is characterized through the way in which claims are associated with various levels of proficiency. Often claims are defined in the form: "Examinee is likely to be successful in tasks with $\overline{\text{characteristic}}$." These characteristics are task model variables.
6. *Spanning the Construct (Content Validity).* Valid definitions of the competency variables require evidence from tasks in a mixture of contexts. Certain task model variables define the key contexts for each task, and assessment assembly rules ensure that an appropriate mix of contexts is put into the final assessment form.

It is important to note that these roles are not mutually exclusive. For example, almost all task model variables support the creation of task variants. Also a variable could easily affect both difficulty and discrimination, and variables that characterize proficiency almost always affect difficulty as well.

The following sections explore each of these roles in greater detail. Drawing from the 2 running examples (i.e., Mathematics Word Problems and Newton's Playground), how each of the roles play out in the different contexts is explored. In particular, the article relates how these task model variables have been identified and explored in the context of the ongoing development of *Newton's Playground.*

CREATING TASK VARIANTS

There are a large number of reasons why assessment designers need variants of tasks. In high-stakes testing situations, task variants guard against memorization of answers—examinees must recognize the deep structure rather than memorize the surface features of the task. Educational researchers often need parallel test forms for use as a pretest and a posttest. In intelligent tutoring systems, the student may need a lot of practice with a particular task type. In the last case, the ability to automatically generate task variants using the computer is very helpful.

Task model variables aid in this process by letting the task authors know what aspects of the task can be changed. In some cases, this can be used generatively, with the task designer deliberately manipulating this feature. An extreme case of this is computer generation wherein the value of the variable is chosen from a specified set of values by a computer, possibly just before the task is presented to the examinee. When presentation material is chosen from authentic material, task model variables help describe what is appropriate. For example, in developing a statistics test, if a task is meant to test knowledge of the *t*-test, then the test developer needs to find a data set for which the *t*-test is an appropriate analysis technique.

When creating task variants, it is often as important to know which task model variables *will not* affect the psychometric properties of a task as to know which ones will. Irvine, Dann, and Anderson (1990) (see also Collis, Tapsfield, Irvine, Dann, & Wright, 1995) introduce the term *radical* to describe features that change the evidentiary properties of the task and the term *incidental* for features that do not. Note that task designers may have theories about which task model variables are radical and which are incidental. It is usually worth testing these theories with members of the target population.

Almost all task model variables have a role in creating task variants, but many of them take on one or more of the roles listed below. It is only the variables that take on no other role that are truly incidentals. The rest of this article explores the possible ways in which task model variables could be a radical.

Math word problems

Consider the "Word Problems Involving Uniform Motion" task model. Each of the blanks is a task model variable. It is relatively easy to see how a computer might generate tasks from this template by randomly selecting from a set of possible values to put into the blanks.

It is somewhat more difficult to figure out which of the possible changes are radical and which are incidental. While changing the numbers in the problem may seem like it will not affect the difficulty, that is only true within a range. Problems with 2-digit numbers are most likely to be easier than problems with 5-digit numbers or fractions or decimals. Changing the numbers to algebraic expressions changes the construct being measured by the task: that is a radical change.

Variables like the names of the actors, the modes of transportation, and the units are mostly incidental, but there are likely limits to how those variables can be changed as well; for example, using unfamiliar units (e.g., furlongs per fortnight as a unit of speed) is likely to cause confusion and create construct irrelevant variance in the measurement, which will weaken the evidentiary strength of the task. If artwork was presented with the task, then changing the mode of transportation beyond a specified set would require new artwork, which would be an additional expense.

In these cases, the need for task variants should be balanced with the cost of producing each variant.

Newton's Playground

Most of the information in the game levels in *Newton's Playground* is found in the initial drawing. The level designer draws a number of objects on the screen and can designate which ones are fixed in place and which will move (obeying Newton's laws and the force of gravity) when the game begins. The designer also specifies the position of the balloon (goal) and the initial position of the ball. Finally, the designer gives a name to the level (which is used in the playground screen where the player selects the levels).

Most of the important task model variables are properties of the drawing, or emerge from the interaction of objects in the initial drawing. Some example task model variables are how many (and how large) are the obstacles on the direct path between the ball and the balloon, are any of those obstacles movable and are there decorative objects placed on the level.

One task model variable relates to the relative position of the ball and the balloon. Whether the ball is to the left or the right of the balloon is incidental: A mirror image of a level will have the same evidentiary properties as the level itself (Figure 4 shows a mirror image of the Spider's Web level, Figure 1). On the other hand, whether the ball is above or below the balloon is radical: If the ball starts above the balloon, then ramp solutions become possible (Figure 5). The horizontal distance between the ball and the balloon is probably partially radical: For certain ranges of distances, it doesn't matter much, but if the ball is very close to the balloon, there is not as much room for drawing new objects.

Some task model variables are difficult to derive from the drawn level and need to be captured in the design process. In particular, the level designer usually has in mind which agents of motion (e.g., lever or pendulum) will likely be used to solve the problem. This is important information for many of the evidence rules and hence needs to be captured.

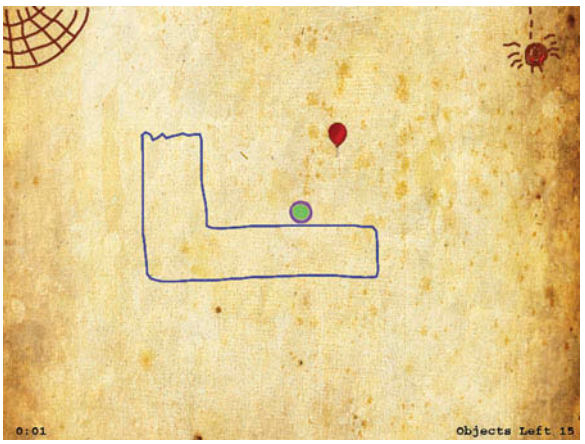

FIGURE 4 Mirror image Spider Web level (color figure available online).

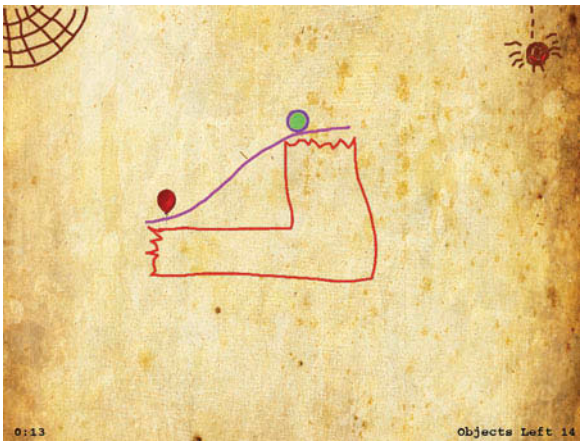

FIGURE 5 Inverted Spider Web level with ramp solution (color figure available online).

Variables that are seemingly incidental may turn out to be radical. Consider the case of objects placed on the screen for decoration, such as the spider web in the corner of the Spider's Web level. Although the intent of the designer was merely to give a theme to the level, it can be used to attach the pivot of a pendulum, thus making pendulum solutions practical in this level. In other cases, decorations can interfere with the motion of the ball or other objects, making the level more difficult.

Note that how a player interacts with problems can also create *emergent task variants* in game-based assessment (Levy, 2012), which does not happen in conventional assessment. Because game-based assessment has some randomness that makes it feel like games, a problem can become 2 different tasks depending on how a player solves the problem. For example, small differences in the length of a pendulum, the mass of its bob, or the timing with which the player draws it can mean that the ball winds up in a different place after being struck, resulting in a new subproblem for the player to solve to reach the goal.

EVIDENTIARY FOCUS: DRIVING DIFFICULTY

Mislevy et al. (2002) identified 1 role for task variables as "mediating the relationship between the observation and the proficiency variable." In item response theory, the relationship between the observable and competency variables are described with an item response function, and Mislevy et al. (2002) are describing variables that change the shape of that function. Measurement theory often discusses 2 ways of changing this function: changing the difficulty (the location of the curve) and changing the discrimination (the slope of the curve). This article breaks the 2 concepts up: This section explores difficulty and the following section, discrimination.

The idea that task features could be used to manipulate the difficulty of a task dates back at least to the linear logistic test model (Fischer, 1973). Fischer modeled the probability of an examinee getting the item correct using a (Rasch) IRT model but replaced the difficulty parameter

with a linear function of a series of task features (task model variables in ECD). If the regression weights for those features were known, then the designers would be able to predict the difficulty of a task from its features.

The subtle differences between the lay and psychometric meanings of "difficult" make the problem of identifying features that might drive difficulty more complex. In the lay definition, a problem (a test item or a game puzzle) is difficult if it takes much effort to solve and if few people are able to solve it. In the psychometric meaning, an observable is difficult if it requires a high level of the measured competency to get "better" levels of the observable. The 2 kinds of difficulty are often, but not always, correlated. To facilitate discussing the relationship between the 2 kinds of difficulty, this article will refer to the lay meaning as *game difficulty,* and *psychometric difficulty* for the other definition.

Mislevy et al. (2002) used the term "Evidentiary Focus" for what we call "Relevant Competencies," but the psychometric difficulty also focuses evidence: changing the difficulty changes the part of the scale for which the task provides the most evidence. A difficult task provides good evidence for the distinction between high and moderate levels of competency while an easy task provides good evidence for the distinction between moderate and low levels of competency. In particular, the highest weight of evidence for distinguishing between 2 levels of competency is achieved when the psychometric difficulty is aligned with the border between the levels.

Math Word Problems

Consider the "Word Problems Involving Uniform Motion" task model. Although switching numbers in that problem makes tasks variants, not all variants will be equal in difficulty. For example, moving from 2-digit numbers to 5-digit numbers makes the problem harder, as does changing from integers to fractions or changing to algebraic expressions.

Often it is not the numbers themselves, but the relationship between them that drives difficulty. For example, if the first and second speeds or distances are the same, then the symmetry makes the problem much easier. On the other hand, if all of the numbers are integers but the solution is not an integer that makes the problem much harder (especially in the constructed-response format), as students expect the answer to be an integer as well (Graf, Harris, Marquez, & Almond, 2008).

Another critical variable arises from the expected solution method. In particular, the *number of sequential steps* needed to solve the problem is a big driver of difficulty. The following sequential steps are usually necessary to successfully solve word-problems based on this task shell: identifying the appropriate word problem type associated with the given problem, combining the relationships among the given known quantities and unknown quantity in a mathematical statement, and solving for the unknown quantity and getting an answer.

However, if the units for the first and second distance are different, then the number of steps changes. The different units add a unit-conversion step. Similarly, if the numbers are fractions but with different denominators, then finding the common denominator becomes an additional step, making the problem harder.

Sometimes task designers will want to deliberately manipulate these factors to increase the difficulty of the task. But at other times they will make the task unsuitable for the purpose of the assessment. For example, task designers would not normally use complex expressions in

the task if the assessment is designed for elementary-age students. Here the change focuses the measurement on the wrong portion of the θ_M scale—one that is outside the expected range for the target population. If the mathematics construct is regarded as multivariate, it changes the variable for which the task provides evidence (This is discussed in the section entitled "Evidentiary Focus: Which Proficiencies").

Similarly, moving to 6-digit integers can make it much more likely for the student to make careless errors. If this is not part of the intended construct, then this can lower the discrimination of the task. Fewer students will get the right answer (especially in the constructed-response format), but that does not mean that the students require higher levels of θ_M to improve their probability.

The distinction between game and psychometric difficulty is important even for conventional-item writing. Consider the translation of the word problem into Sanskrit. This will make the problem more difficult (except for persons with some fluency in Sanskrit) in the sense that fewer people will get the correct answer. On the other hand, this increase in difficulty will have very little relationship to the constructs of interest. This is an example of game difficulty, and not psychometric difficulty.

Newton's Playground

Because there are 3 target competency variables in *Newton's Playground,* there are 3 different psychometric difficulties: 1 for physics understanding, 1 for persistence, and 1 for creativity. Furthermore, the difficulty of individual observables might be different. For example, if a ramp solution is obvious and a springboard solution is not, then the springboard solution might require more knowledge of physics (or more creativity) than the ramp solution.

It is important not to confuse game difficulty and psychometric difficulty during the design phase. For example, Figure 6 is a problem called Around That Tree. It is a fairly easy problem because all a player needs to do is to draw a simple ramp that guides the ball to the balloon.

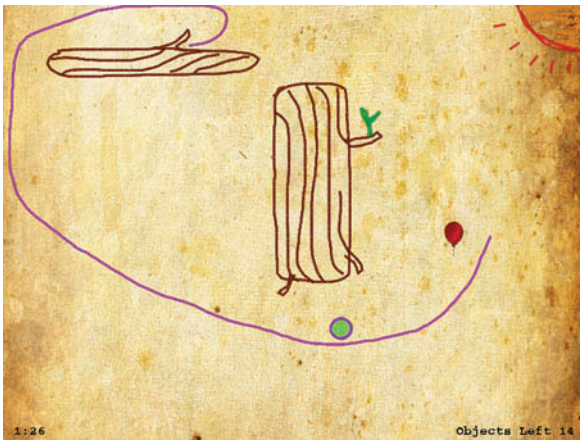

FIGURE 6 Around That Tree game level with ramp solution (color figure available online).

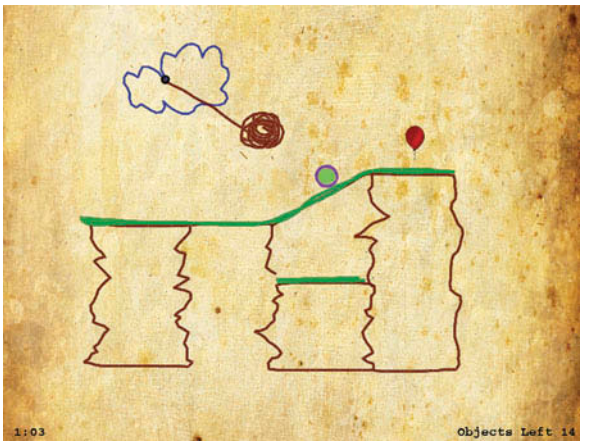

FIGURE 7 Golfing game level with pendulum solution (color figure available online).

However, because it is an easy problem with a limited number of possible solutions, nonramp solutions are rare. Only players with high creativity will think of nonramp solutions. Therefore, this level yields relatively weak evidence in terms of physics, yet provides strong evidence regarding creativity if a nonramp solution is used.

Another illustration of this idea is the game level called Golfing (Figure 7). The name of this level provides a clue that the solution is to draw a pendulum (a golf club) which will drive the ball to the balloon. This makes the problem easier for testing understanding of angular momentum (as we have eliminated the need to recognize pendulum as a solution path), but it makes it harder for measuring creativity as it is unlikely that players will try nonpendulum solutions. In other words, this level is only good for measuring creativity at the high end of the scale.

Thus game difficulty is an important task model variable, one that influences the psychometric difficulty of many of the relationships between the observables and the competency variables. However, the nature of that relationship is different for each pair. For physics understanding, we predict that game difficulty will have a monotonic relationship with psychometric difficulty: harder problems will require more understanding of physics to solve. For creativity, we expect the relationship to be U-shaped. For easy problems, there will be an obvious solution and it will take a great deal of creativity to generate original solutions. For hard problems, there will be a very limited number of possible solutions, and it will take a great deal of creativity to fluidly generate a number of solutions. We predict that problems of moderate game difficulty will be the easiest in which to express creativity.

The relationship between game difficulty and persistence is more complex. For people who cannot solve easy problems, it is probably quite difficult to persist given that the player is "just not getting it." The problem is that most people will recognize the solution quickly, and thus most people will not need much persistence. Thus, it is not difficulty (high levels of persistence needed to solve the problem) but evidential strength (there is a low correlation between solution time and persistence) that task difficulty affects. The following section explores evidential strength.

Finally, it is important to note that these are theories about difficulty not facts. Often it is the case that the theories of difficulty are derived from the cognitive theory about the domains. Pilot testing, pretesting, and play testing are important to validate the theories. Observing members of the target population playing with the game is important for validating the designers' theories of difficulties. Problems may be unexpectedly easy or difficult, and there may be game features that drive difficulty that were not identified in the cognitive model.

EVIDENTIAL STRENGTH (DISCRIMINATION)

Some kinds of observations provide stronger evidence than other kinds. Good (1985) (see also Schum, 1994) defines the *weight of evidence* as follows. Let H be a hypothesis (set of claims) whose truth is to be established (a typical example is that a student falls into a particular category or range on a competency variable), and let $\bar{H}$ represent the proposition that the hypothesis does not hold. Let E be a bit of evidence (for example the fact that an observable value took on a certain value after a student attempted a certain task. The weight of evidence is

$$W(H:E) = \log \frac{\Pr(E|H)}{\Pr\left(E|\bar{H}\right)} = \log \frac{\Pr(H|E)}{\Pr\left(\bar{H}|E\right)} - \log \frac{\Pr(H)}{\Pr\left(\bar{H}\right)}. \quad (1)$$

The right hand side of that equation comes from applying Bayes theorem in log odds form: The weight of evidence is the difference between the prior log odds and the posterior log odds.

The importance of this formal definition in task design is that a task has good evidentiary strength for measuring H when $\Pr(E|H)$ is large and $\Pr\left(E|\bar{H}\right)$ is small. In other words, people for whom the target claims do hold produce very different work products from people for whom the claims do not hold. As various task features will control how various classes of people are likely to respond, identifying these task variables—the ones that affect weight of evidence—facilitates writing better tasks.

If H corresponds to the upper portion of a unidimensional scale and $\bar{H}$ corresponds to the lower portion, then we can relate evidentiary strength to the IRT difficulty and discrimination parameters. If the task is very easy (in the psychometric sense), then both $\Pr(E|H)$ and $\Pr\left(E|\bar{H}\right)$ will be large, so the weight of evidence is low. Similarly, if the task is hard, then both $\Pr(E|H)$ and $\Pr\left(E|\bar{H}\right)$ will be low, so again the weight of evidence will be low. Thus, targeting the difficulty of the task to be close to the borderline of H maximizes the strength of evidence. Section 6 discusses ways to manipulate difficulty.

The other IRT parameter that affects weight of evidence is discrimination. If the slope of the item response function is steep near the cut point between $\bar{H}$ and$H$, then the difference between $\Pr(E|H)$ and $\Pr\left(E|\bar{H}\right)$ will be large; if the slope is not steep, the weight of evidence will be small. Thus, improving the discrimination improves the weight of evidence.

One kind of variable that can affect the weight of evidence is the requirement of nonfocal knowledge, skills, and abilities (Hansen & Mislevy, 2004). The distribution of that nonfocal knowledge in the population can cause variance in the outcome variable unrelated to the claims targeted by the task. Consider once again a question about the t-test on a statistics exam that uses authentic data. Suppose that the data set uses a medical experiment as an example. This would require the student to read and understand (at least at a shallow level) medical vocabulary. If this

task was used in a place where most students are unfamiliar with that vocabulary (say a college of education), then (lack of) knowledge of medical vocabulary (a construct-irrelevant skill) would depress both $\Pr(E|H)$ and $\Pr\left(E|\bar{H}\right)$, lowering the weight of evidence.

It is important to distinguish between task variables and person variables. In the statistics example, the context of the problem is the task variable and familiarity with the context is the person variable. Cases in which the person variable corresponding to the task variable is not evenly distributed throughout the population threaten the fairness of the test. Consider once more the use of a medical data set on a statistics test. If some of the students are medical students and others are from other disciplines, the context of the problem means that the weight of evidence is different for the 2 groups of students. Similarly, if some of the students are foreign students who do not have extensive knowledge of English outside of their discipline, they are likely to find the problem more difficult. Restricting the values of the context variable to values that all students are likely to be equally familiar with (all familiar or all unfamiliar) helps eliminate the dependence on a nonfocal ability.

On the other hand, often we are willing to accept tasks that have lower evidentiary strength because they tap parts of the construct that would be otherwise difficult to assess. For example, removing all context from the statistics problem would eliminate much of the construct irrelevant variance but it would not also allow the designers to assess the examinees' ability to transfer their statistics knowledge to new contexts.

Math Word Problems

Consider the "Word Problems Involving Uniform Motion" task model. Here the vocabulary used in the stem affects discrimination. Say that the task author created an instance of the uniform-motion problem using sailing as the example and then used nautical jargon for the words of motion (e.g., "tacking" and "reaching"). The use of unfamiliar language would make the task harder for students with strong math skills and students with weak math skills. Similarly, using unfamiliar units for distance and speed (e.g., furlongs per fortnight) will diminish the evidentiary strength of the problem by increasing the dependence on a nonfocal ability (vocabulary).

As already discussed, using an excessive number of digits depresses both $\Pr(E|H)$ and $\Pr\left(E|\bar{H}\right)$. Thus it lowers evidentiary strength.

It is actually quite easy to lower the evidentiary strength of a task. Introducing any nonfocal knowledge, skill, or ability will have that effect. Raising the evidentiary strength is harder, it involves searching for and eliminating these nonfocal elements. One needs to be careful. In the word problem domain, vocabulary is mostly nonfocal, except that a critical part of the claims made by the assessment is that the students can recognize the schemas from natural language. So a certain amount of vocabulary is necessary to support that claim.

Newton's Playground

Game difficulty can affect evidential strength as well as psychometric difficulty. Consider 2 problems, Golfing (Figure 8) and Smiley (Figure 9). Both problems are designed to assess the player's understanding of how A pendulum works (i.e., angular momentum). For Golfing, the player needs to draw a golf-club-like pendulum on the cloud to swing the ball to the balloon. There is a cloud

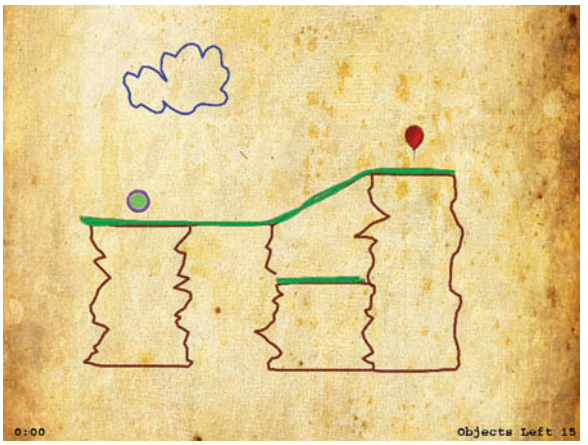

FIGURE 8 Golfing level: Strong evidence for angular momentum (color figure available online).

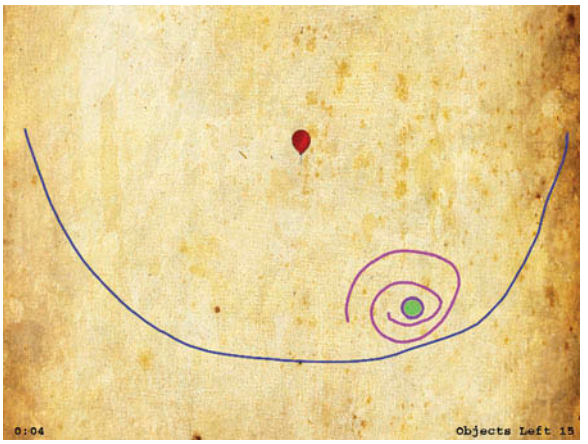

FIGURE 9 Smiley level: Weaker evidence for angular momentum (color figure available online).

right above the ball hinting at where the pivot needs to go, and the name of the problem "Golfing" provides a hint to use a pendulum. For Smiley, the optimal placement of the pivot is along the side of the main circle, neither the name of the level nor the drawing provide a clue that a pendulum is a good solution. Smiley is harder than Golfing for both players with strong and with weak understanding of angular momentum. This means the difference between $\Pr(E|H)$ and $\Pr\left(E|\bar{H}\right)$ is larger for Golfing and that level has more evidential strength. Smiley provides a reason other than poor understanding of pendulums that a player might not attempt a pendulum solution: the player may simply not have realized that it can be solved by a pendulum. Therefore, the task

model variable here that changes the evidential strength is *how obvious it is to players where to position the pivot of the pendulum.*

While Smiley provides less evidence for physics understanding than Golfing, it provides more evidence for persistence. As Golfing is relatively easy, it usually does not provide much evidence for persistence. Most players solve the level quickly. Smiley, which requires both identifying the pendulum solution and fine tuning the pivot placement, requires more persistence and hence provides better evidence for the persistence variable.

The situation with the creativity competency is more complex. Most of the time, players will use the pendulum solution for Golfing. This is not particularly original, thus, this is not strong evidence for originality. Occasionally, a player may attempt a solution using another agent. This is highly original and produces strong evidence for creativity. Because the pendulum solution is much more likely, the expected weight of evidence of Golfing for creativity is still low.

Note that for assessments embedded within games, game familiarity is an obvious nonfocal ability. It is also a troublesome one because 1 subpopulation (e.g., the gamers) will have more game familiarity than another. *Newton's Playground* is a casual puzzle game and as such is less likely to produce this kind of problem than other game genres. Still, this is a potential source of unfairness that should be studied. Note that conventional tests (especially multiple choice) are facilitated by a similar test familiarity skill, which weakens their evidential strength.

EVIDENTIARY FOCUS: WHICH PROFICIENCIES?

When the competency model has more than 1 variable, this adds another dimension to the problem of determining evidentiary focus of a task. In particular, task designers need to determine which competency variables are relevant for each particular task. Changing the task features should allow task designers to produce a mix of tasks with different evidentiary foci.

Note that in the case of a unidimensional assessment, this reduces to the previous problem of determining evidentiary strength (previous section). In this case, any additional competency variable that influences the probability of observing certain values for the observable variables will be construct irrelevant. That means it will reduce the evidentiary impact.

Tatsuoka (1983) recognized the duality between the task features that formed the columns of the Q-matrix and the claims that we might wish to make about students. Different tasks will load on different proficiency variables, based on their task model variables. The Q-matrix emerges from these loadings and the collection of tasks defined by the assembly model.

Typically, assessment designers need tasks that tap different patterns of competencies in order to be able to identify various competency profiles. Leighton and Gierl (2007) present some methods for manipulating a Q-matrix to determine which patterns can be readily identified. Changing the Q-matrix to produce better patterns of evidence requires that task designers be able to manipulate which competency variables are relevant for each task.

Math Word Problems

Consider once again the "Word Problems Involving Uniform Motion" task model, but this time instead of the 2 variables competency model, suppose that there are 5 mathematical strands

following the framework laid out in the National Assessment for Educational Progress (NAEP; NAGB, 2006).

Consider a task model variable that describes the domain of values chosen for distances or speeds. If the choices are restricted to integers, or rational numbers, then the task will provide evidence for the Numbers and Operations competency variable. If the choices are algebraic expressions, then the task will provide evidence for the Algebra and Functions competency. Similarly, if the Numbers and Operations competency variable were further divided, then the choice of integers or fractions for the values for distances and speeds would focus on different competencies.

The alternative to the multivariate view is to regard the various strands of the NAEP framework as different regions of an underlying mathematics competency. The question that assessment designers face is then, are the claims targeted by the task consistent with the purpose of the assessment or are they outside the scope. For example, it is probably not appropriate to give a task involving algebra to 4th-grade students. On the other hand, we expect 4th-grade students to be able to manipulate both integers and fractions, so we probably need to ensure that problems using both values are represented in the assessment (This is discussed in the section entitled "Spanning the Construct").

A consequence of this focus on unidimensional assessments is that this role of task model variables has often been overlooked by designers. In the unidimensional case, rather than focusing on which competencies the task provides evidence for, the question becomes, "Does the task provide evidence for the target competency and claims?" If the answer is yes and the evidence is sufficiently strong, the task is used in the assessment; if not, the task is rejected as inappropriate.

Newton's Playground

In the *Newton's Playground* competency model, the physics-understanding competency is split into 2 subcompetencies[2]: understanding potential and kinetic energy and understanding angular momentum. While the first is required to a certain extent for almost all *Newton's Playground* problems, the latter comes into play only when the player uses a lever, a springboard or a pendulum.

Thus problems that permit ramp solutions involve different combinations of competency variables (or else the way that the tasks combine the proficiencies is different—a disjunctive or compensatory relationship rather than a conjunctive relationship). As ramp solutions are very difficult when the balloon is higher than the ball, the relative height of ball and balloon is a simple variable that controls which of the 2 subcompetencies come into play.

For example, Figure 10 is a problem called Shark. This problem can be solved by drawing a lever on the top of the shark's fin but is impossible to solve using a ramp. Thus the focus of this task includes both angular momentum and potential/kinetic energy.

The ultimate balance of which proficiencies are measured is controlled by the assembly model. The task model variables help identify which tasks are best at assessing which aspects of proficiency. For example, tasks with a constrained solution path are better for assessing physics proficiency, while tasks with a less constrained solution path are better for assessing creativity (as players can explore various solution options). The assembly model addresses how much of each

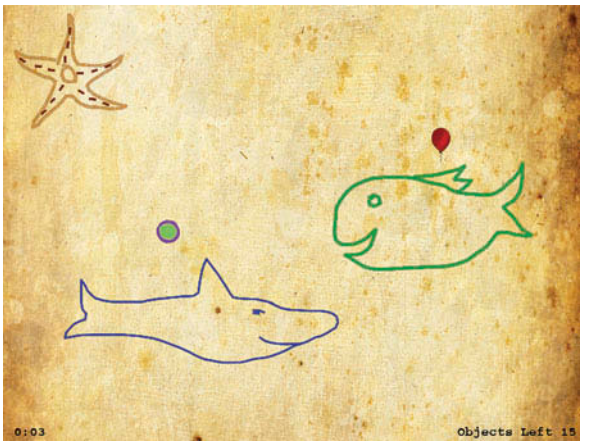

FIGURE 10 Shark: Level provides no information about ramps (color figure available online).

is used, either by balancing how many tasks of each type are used or by specifying how much information about each proficiency should be gathered.

CHARACTERIZING PROFICIENCY

In ECD, competency variables are defined through the claims that they support. Those claims are often related to the kinds of tasks that an examinee at a particular level of competency should be able to perform. The American Council on the Teaching of Foreign Languages (ACTFL; 1989) scales for spoken language proficiency are prime examples of this. At progressively higher levels of language ability the examinee is able to carry on conversations about progressively more complex and abstract subjects. Here the complexity of the subject to be discussed is a task model variable whose values characterize a level of the ACTFL scale.

This role of the task model variable is closely related to its role in determining psychometric difficulty (See the section entitled "Evidentiary Focus: Driving Diffculty"). If the competency variables are modeled using the Rasch model, the Wright map and its variants (Wilson, 2005) provide a mechanism for putting the items onto a scale. Beaton and Allen (1992) show how scale maps can be created for other response modeling frameworks by looking at the point at which a certain percentage of examinees are able to achieve that level of performance. Looking at the task characteristics of tasks that cluster on similar parts of the scale provides an easy-to-understand description of what that part of the scale means.

When the scales are ordered categories, characterizing the scale can still be done using tasks. As the examinee passes to the next higher level of competency, there should be some new tasks that the examinee is usually able to perform (or new observable features that are usually apparent in the examinee's work). Looking at the task features for the tasks that are in this difference set provides information about the operational difference between the 2 levels of competency.

Math Word Problems

Suppose that the word problems are being used to assess a student's level of algebraic thinking. As the student develops more algebra skills, it follows that the student is able to solve problems in which the unknown quantities are increasingly abstract. The *level of abstraction* of the problem is a task model variable that can be used to characterize proficiency. Figure 11 shows a construct map made with the proficiency levels on the left-hand side and the task descriptions on the right. The figure is meant to suggest a scale and its implicit claims. People who are at a given position on the left-hand side of the scale can do the tasks that are at that level or lower on the right-hand side.

Another task model variable that could be used to characterize proficiency is the *number of steps* required; students at higher levels of proficiency are able to handle multistep problems. Also, specific kinds of steps might be associated with various stages of development—for example, problems involving unit conversions might be associated with a specific level on the mathematics construct.

Newton's Playground

Finding variables that characterize proficiency in *Newton's Playground* is a bit harder. In particular, most of the time proficiency is characterized by the features of the work product that the player produces. Some examples only have 2 levels (and hence don't make an interesting map). For example, *understands angular momentum* is one facet of the *conceptual physics* competency.

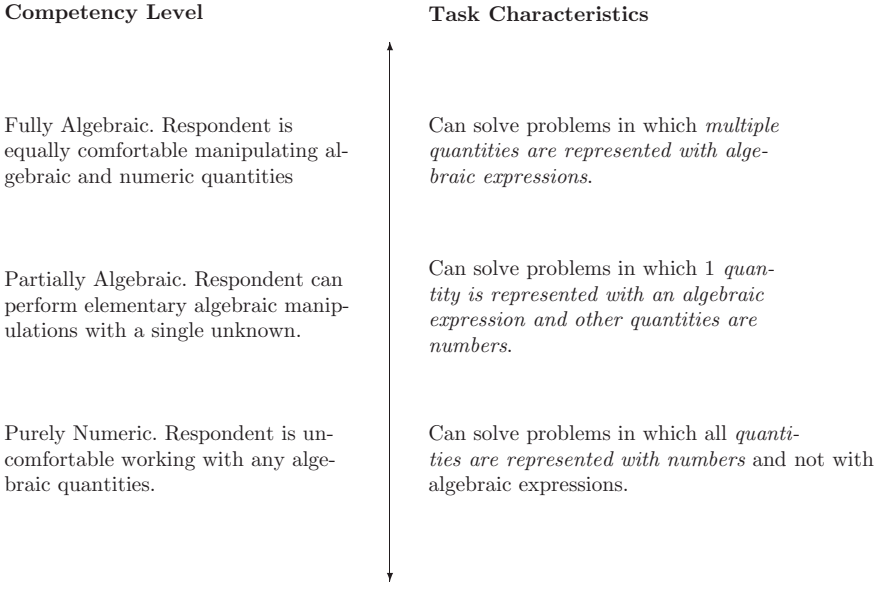

FIGURE 11 Construct map for mathematical problem solving.

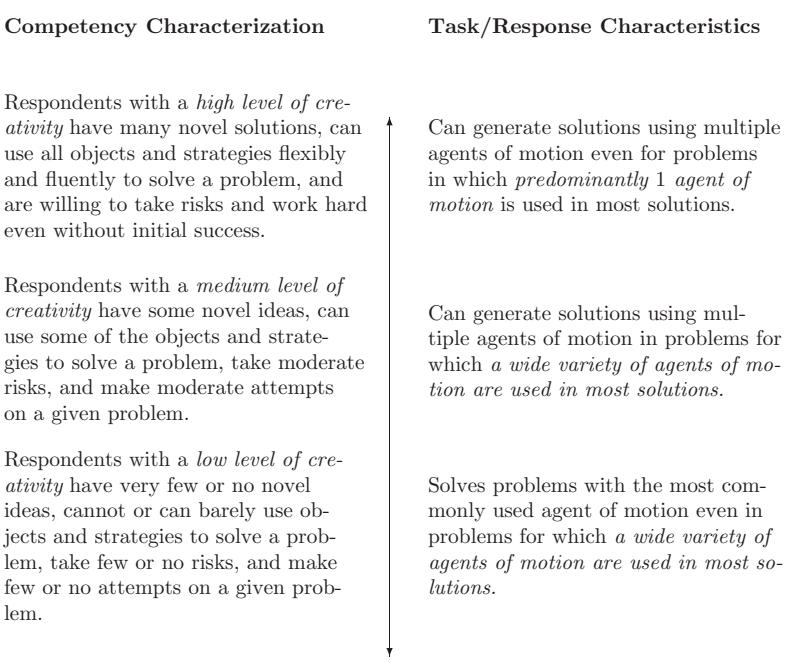

FIGURE 12 Construct map for creativity in *Newton's Playground.*

Here we would expect a person who understands angular momentum to be able to solve problems that require levers, springboards, and pendulums—all of which require knowledge of angular momentum to set up properly.

A more complex example comes from looking at the *originality* facet of the *creativity* competency. Here we expect that a person who is highly creative will come up with solutions that are different from the usual solutions. One way that a solution could be different would be to use a different agent of motion, say to use a springboard when most other solutions use a lever. Here the frequency of springboard solutions is a task model variable (albeit one whose value can only be established from a large sample of solutions to the task). Figure 12 is an example of a creativity-construct map for *Newton's Playground* based on originality of the agent of motion.

Obviously, this is not a complete definition of *creativity.* Figure 12 would need to be put together with a number of other construct maps (some written in terms of observations rather than task descriptions). These in turn would need to be validated by scaling the observable outcomes from the task. Still, maps like this play an important role in associating meaning with latent scales, and to the extent that the context of the task is important and task model variables describe that context, they play a role in establishing the scale as well.

SPANNING THE CONSTRUCT

ECD defines the construct to be measured in terms of the claims about the examinee that the assessment results must support. The challenge in defining an assessment is ensuring that there

is enough support for each of those claims. In particular, there must be enough tasks that directly support each claim. This is controlled through the interaction of the test-assembly rules and the task model variables. Task model variables describe the critical content of each task, and the assembly rules specify minimum and maximum numbers for the corresponding values.

This is closely related to the concept of construct validity (Embretson, 1983; Kane, 2006). In particular, the test design must cover the content, or at least sample from the targeted content in a way that will cover it all with reasonable probability. Again the task model variables describe what part of the content is addressed by each task, and then the test-assembly rules ensure that the distribution is reasonable.

In addition to ensuring that content requirements are met, there are 2 other ways in which the task model variables and the assembly rules interact to improve the psychometric validity of an assessment. First, if there are possible nonfocal knowledge, skills, and abilities that are needed for answering certain items, these can be noted in task model variables. The assembly rules would then specify a maximum for each one. For example, consider a reading assessment for students seeking to enter graduate study. The topics for the reading passages could be chosen from a list of topics (e.g., natural sciences, social sciences, and humanities). The assembly rules would restrict the maximum number of passages for each topic so that no one discipline is overrepresented.

Math Word Problems

The schema theory of Marshall et al. (1987) partitioned the various schemas used to solve word problems to create 5 types: *change, combine, compare, vary and transform*. A reasonable definition of the mathematics construct spans all 5 types, so the test-assembly rules should enforce this constraint.

To do this we set up a series of task model variables representing the schema type. As 1 task model could have multiple schema types (i.e., a compound problem requiring multiple steps), we represent the schema type with 5 binary variables, which take the value of 1 if a subproblem has that schema type and 0 if it does not. The assembly model then has rules requiring minimum numbers of tasks with each of these task model variables set.

Another problem that can arise in the word-problem world (especially if the items are auto-generated) is that the tasks are too similar and that practice with one task will produce learning on the test. In that case we might define similarity through the number of task model variables they have in common, and the assembly rules would restrict tasks that are too similar from appearing on the same form. Note that there are some situations (particularly tutoring contexts) where the repetition and learning of the schema are beneficial, and hence they would not require such a restriction.

Newton's Playground

The *Conceptual Physics* construct in *Newton's Playground* spans all of the agents of motion supported by the game engine: ramps, springboards, levers, and pendulums. Thus, a reasonable assessment of a player's conceptual physics would include the player attempting to solve the problems using all 4 agents.

The task designers cannot control which agent a player uses to solve a problem, but they can make certain solution types that are easier and more obvious than others. The set of levels developed for *Newton's Playground* include a number of tasks for which the designers intend that the problem be solved with 1 agent of motion (e.g., a lever problem). In these 1-agent problems, solutions with the target agent are relatively obvious and solutions with nontarget agents are less obvious. A number of these were developed for each of the 4 agents of motion.

Again we cannot force a player to play a particular level, but the levels are grouped into "playgrounds" with the suggestion that a player work on 1 or 2 playgrounds during a typical playing session. The playgrounds are balanced much like the forms on a traditional fixed-form test. In this case, each form includes some problems that target each of the 4 agents as well as some more-open-ended problems for assessing creativity.

Newton's Playground is a game consisting of a series of discrete levels, which makes the assembly closer to traditional form design. In a more-open-ended game, other mechanisms would be needed to ensure that the task attempted by the player spans the intended construct. One solution would be to offer greater in-game rewards for attempting tasks that cover parts of the construct that have not yet be adequately measured.

SOME NOTES ON APPLYING THESE IDEAS

Mislevy, Behrens, DiCerbo, Frezzo, and West (2012) discuss issues involved with incorporating assessments into games and simulations. Their first and most important point is that assessment is not just about scores but also about the logical argument that is built up around those scores. Consider an assessment that consists merely of questions taken from a book of puzzles. Scoring high on this puzzle assessment only supports the claim that the examinee is good at puzzles. Linking the competencies needed to solve the puzzles to claims about the examinee transforms the score from a mere datum to evidence about those claims.

Test designers are seldom satisfied with measuring a single claim, but usually wish to draw inferences about a collection of claims organized into 1 or more scales. In designing the assessment, the authors must adjust the focus of the tasks to measure claims on different parts of the scale. When there are multiple scales, the authors must also determine which competency variable(s) the observable variables in the task load on.

The variables that will allow the test designers to adjust the focus of evidence are specific to the types of tasks and the assessment purpose. Identifying them in general is an impossible task. However, the practice of thinking about the features of tasks that can be varied and recording those features and those roles helps on many levels. First, it provides a framework for test designers' introspection about the task design and for discussions among test designers about what elements of the task are important. Second, it provides specification, which helps train new task authors (Riconscente et al., 2005). Third, it provides a design information for computer scientists to automate the production of task variants (Deane, Graf, Higgins, Futagi, & Lawless, 2006).

The distinction between game and psychometric difficulty is particularly useful, for the construction of both conventional and game-based tasks. In conventional assessment, the authors usually want to minimize game difficulty (and the associated construct-irrelevant variation) while manipulating task difficulty to meet the information targets of the assessment. The authors of game-based assessment must also manipulate the combined game and psychometric difficulty to

make the game optimally engaging. However, the more that engagement can be supported with psychometric difficulty, the better: game levels that have too much nonpsychometric difficulty will provide weak evidence.

This article focuses on the roles of task model variables that task authors can consider in the process of designing assessments. For interactive assessment such as games and simulations, however, it is unlikely that even task authors can fully predetermine (or predict) all possible ways examinees interact with the system (Levy, 2012). By their very nature, games and simulations provide examinees with more opportunities to choose and make actions. Therefore, in the process of developing interactive assessment, the knowledge in the task model can be used to build agents that recognize evidence-bearing patterns that emerge from the examinee's interactions with the assessment system.

One challenge that arises from this player freedom is covering the construct. In various trials of *Newton's Playground*, most players followed the sequence of levels provided by the system, but many did not. For the players that skip around, or do not complete much of the game, how can the designers ensure adequate coverage of all of the constructs to be measured? There may be clever solutions that involve manipulating game mechanisms to encourage players to provide information about competencies not adequately covered. Even so, the game designers need to analyze the level-completion data to make sure that adequate coverage is achieved.

Another intriguing possibility offered by game and simulation-based assessments is the capability of tracking player progress toward a solution. In particular, specific kinds of events within a solution may provide evidence of a proficiency. For example, in *Newton's Playground* any time the player draws a lever, that provides some evidence that the player understands levers. There are two challenges in using this kind of information: (1) the level designers must identify useful observations they can make about a solution or attempted solution and (2) the observations based on partial solution are likely statistically dependent. This capability is the subject of ongoing and future research (Kerr & Chung, 2013). As we learn more about capturing solution-path information, we expect that test developers will fold those lessons back into task design.

At first, there will only be theories about which task features drive the psychometric properties. These theories will require testing (possibly of the kind described in Lawless, Sabatini, Deane, Bejar, & Chen, 2012). This will inevitably lead to refinement of the theories. Eventually, test developers will be able to use the task features to reduce the amount of pretesting required for new tasks, or even to avoid pretesting altogether in low-stakes situations (Mislevy, Sheehan, & Wingersky, 1993).

NOTES

1. Based on our experience using the Mislevy et al. (2002) paper with students, we have given the roles simpler names. Also, the 3 roles starting with "focusing evidence" were originally contained in 2 roles :1 "focusing evidence" and "mediating the relationship between observation and proficiency."
2. The domain of physics understanding involves many more subcompetencies, but 2 are sufficient to cover the range of problems possible within the scope of the *Newton's Playground* simulation.

ACKNOWLEDGMENTS

I would like to acknowledge the students in the Fall 2012 section of the Florida State University class EDF 5448 (Scale and Instrument Development): Jackie Cocke, Urska Dobersek, Antonietta Echezula, Gonca Gul, Yuhua Guo, Herlanda Hampton, Deanna Harris-McKoy, Bi-Jen Hsieh, Karin Jeffrey, Jiwon Nam, Christine Ouma, Umit Tokac, and Xinrong Xue, whose classroom discussion of these issues helped us clarify many of them. Bob Mislevy provided a number of extremely helpful comments on a draft of the manuscript.

Many aspects of the *Newton's Playground* examples are based on work of the *Newton's Playground* team Val Shute, PI. In addition to the authors, the team includes Matthew Ventura, Matthew Small, Don Franceschetti, Lubin Wang, and Weinan Zhao.

FUNDING

Work on *Newton's Playground* and this article was supported by the Bill and Melinda Gates Foundation, U.S. Programs Grant No. #0PP1035331, *Games as Learning/Assessment: Stealth Assessment.* Any opinions expressed are solely those of the authors.

REFERENCES

Almond, R. G. (2010). "I can name that Bayesian network in two matrixes." *International Journal of Approximate Reasoning*, *51*, 167–178. doi: 10.1016/j.ijar.2009.04.005

Almond, R. G., DiBello, L., Jenkins, F., Mislevy, R. J., Senturk, D., Steinberg, L. S., & Yan, D. (2001). Models for conditional probability tables in educational assessment. In T. Jaakkola & T. Richardson (Eds.), *Artificial intelligence and statistics 2001* (p. 137–143). San Francisco, CA: Morgan Kaufmann.

Almond, R. G., Kim, Y. J., Shute, V. J., & Ventura, M. (2013). Debugging the evidence chain. In R. G. Almond & O. Mengshoel (Eds.), *Proceedings of the 2013 UAI application workshops: Big data meet complex models and models for spatial, temporal and network data* (uai2013aw) (pp. 1–10). Retrieved from http://ceur-ws.org/Vol-1024/paper-01.pdf

American Council on the Teaching of Foreign Languages (Ed.). (1989). *ACTFL proficiency guidelines*. Alexandria, VA: American Council on the Teaching of Foreign Languages.

Bart, W. M., Post, T., Behr, M. J., & Lesh, R. (1994). A diagnostic analysis of a proportional reasoning test item: An introduction to the properties of a semi-dense item. *Focus on Learning Problems in Mathematics*, *16* (3), 1–11.

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, *17*(2), 192–204.

Catto, E. (2011). Box2D v2.2.0 user manual [Computer software manual]. Retrieved from http://box2d.org/

Collis, J. M., Tapsfield, P. G. C., Irvine, S. H., Dann, P. L., & Wright, D. (1995). The British Army Recruit Battery goes operational: From theory to practice in computer-based testing using item generation techniques. *International Journal of Selection and Assessment*, *3*, 96–104.

Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology*, *79*(4), 347–362.

Deane, P., Graf, E. A., Higgins, D., Futagi, Y., & Lawless, R. (2006). *Model analysis and model creation: Capturing the task-model structure of quantitative item domains* (Research Report No. RR-06-11). Educational Testing Service. Retrieved from http://www.ets.org/research/policyj_researchj_reports/rr-06-11

Dori, Y. J. (2003). From nationwide standardized testing to school-based alternative embedded assessment in Israel: Student' performance in the matriculation 2000 project. *Journal of Research in Science Teaching*, *40*(1), 34–52.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–197.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.

Gee, J. P. (2010). Human action and social groups as the natural home of assessment: Thoughts on 21st century learning and assessment. In V. J. Shute & B. J. Becker (Eds.), *Innovative assessment for the 21st century: Supporting educational needs* (pp. 13–40). New York, NY: Springer.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15* (1), 1–38.

Good, I. J. (1985). Weight of evidence: A brief survey. In J. Bernardo, M. DeGroot, D. Lindley, & A. Smith (Eds.), *Bayesian statistics 2* (pp. 249–269). Amsterdam, The Netherlands: Elseveir Science Publisher.

Graf, E. A. (2008). *Approaches to the design of diagnostic item models* (Research Report No. RR-08-07). Educational Testing Service. Retrieved from http://www.ets.org/research/researcher/RR-08-07.html

Graf, E. A., Harris, K., Marquez, E., & Almond, R. G. (2008, March). *A cognitively based assessment system for mathematics competency*. Paper presented at the Annual Meeting of the American Educational Research Association (AERA), New York, NY.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals ofitem response theory*. Newbury Park, CA: Sage.

Hansen, E. G., & Mislevy, R. J. (2004). Toward a unified validity framework for ensuring access to assessments by individuals with disabilities and English language learners. In *Paper presented at the Annual Meeting of the national council on measurement in education (NCME)*, San Diego, CA. Retrieved from http://www.ets.org/research/dload/NCME2004-Hansen.pdf

Hiebert, J., & Wearne, D. (1993). Instructional task, classroom discourse, and students' learning in second grade. *American Educational Research Journal*, *30*, 393–425.

Irvine, S. H., Dann, P. L., & Anderson, J. D. (1990). Towards a theory of algorithm-determined cognitive test construction. *British Journal of Psychology*, *81*, 173–195.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17–64). American Council on Education/Praeger.

Kerr, D., & Chung, G. K. (2013). Identifying learning trajectories in an educational video game. In R. G. Almond & O. Mengshoel (Eds.), *Proceedings of the 2013 UAI application workshops: Big data meet complex models and models for spatial, temporal and network data* (uai2013aw) (pp. 20–28). Retrieved from http://ceur-ws.org/Vol-1024/paper-03.pdf

Lawless, R. R., Sabatini, J. P., Deane, P., Bejar, I. I., & Chen, L. (2012). *Measuring the depth of semantic knolwedge in academic and domain-specific vocabulary*. Paper presented at the 2012 Annual Meeting of the American Educational Research Association (AERA), Vancouver, Canada.

Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment: Theories and applications*. New York, NY: Cambridge University Press.

Levy, R. (2012). *Psychometric advances, opportunities, and challenges for simulation-based assessment* (Center for K-12 Assessment and Performance Management Report). Educational Testing Service.

Marshall, S., Pribe, C., & Smith, J. (1987). *Schema knowledge structures for representing and understanding arithmetic story problems* (Technical Report No. Contract No. N00014-85-K-0661). Office of Naval Research.

Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review*, *78*(2), 333–368.

Mislevy, R. J., Behrens, J. T., DiCerbo, K. E., Frezzo, D. C., & West, P. (2012). Three things game designers need to know about assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives* (pp. 59–81). Springer.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. S. (1993). How to equate tests with little or no data. *Journal of Educational Measurement*, *30*, 55–78.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the roles of task model variables in assessment design. In S. H. Irvine & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and practice* (pp. 97–128). Erlbaum.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment (with discussion). *Measurement: Interdisciplinary Research and Perspective*, *1* (1), 3–62.

Mislevy, R. J., Steinberg, L. S., Almond, R. G., & Lukas, J. F. (2006). Concepts, terminology and basic models of evidence-centered design. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 15–47). Mahwah, NJ: Lawrence Erlbaum Associates.

Morales, R. V., Shute, V. J., & Pellegrino, J. W. (1985). Developmental differences in understanding and solving simple word problems. *Cognition and Instruction*, *2*(1), 41–57.

National Assessment Governing Board (NAGB). (2006). *Mathematics framework for the 2007 National Assessment of Educational Progress*. U.S. Department of Education. Retrieved from http://www.nagb.org/frameworks/math_07.pdf

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics.* Washington, DC: Authors. Retrieved December, 2012 from http://www.commoncorestandards.org/

National Council of Teachers of Mathematics. (1989). *Curriculum and performance standards for school mathematics*. Retrieved from http://www.mathcurriculumcenter.org/PDFS/CCM/summaries/standards_summary.pdf

Riconscente, M., Mislevy, R. J., & Hamel, L. (2005). *An introduction to PADI task templates* (PADI Technical Report No. 3). SRI International. Retrieved from http://padi.sri.com/downloads/TR3_Templates.pdf

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

Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. Wiley.

Shute, V. J., Hansen, E. G., & Almond, R. G. (2008). You can't fatten a hog by weighing it—or can you? Evaluating an assessment for learning system called ACED. *International Journal of Artificial Intelligence in Education*, *18*(4), 289–316.

Shute, V. J., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. MIT series.

Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. J. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). Routledge, Taylor and Francis.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York, NY: Springer-Verlag.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment*, *11*(1), 65–83.