Chapter 10
Driving Assessment of Students' Explanations in Game Dialog Using Computer-Adaptive Testing and Hidden Markov Modeling

Douglas B. Clark, Mario M. Martinez-Garza, Gautam Biswas, Richard M. Luecht, and Pratim Sengupta

10.1 Introduction

The three central components of science education in the classroom—learning, teaching, and assessments—have traditionally focused on facts and rote learning. Students in most science classrooms have traditionally memorized equations and names of chemical elements, cloud types, bones, and organs and are usually not provided with meaningful opportunities to develop deep understandings of the relevant phenomena, or use such knowledge to explore natural phenomena. These traditional approaches to learning, teaching, and assessment, however, do not align with current goals for science literacy that focus on students' ability to engage in extended problem solving that involves exploration, explanation, application of integrated conceptual knowledge to rich and realistic contexts (AAAS, 1993; NRC,

D.B. Clark (✉) • M.M. Martinez-Garza
Learning, Environment, and Design Lab, Peabody College, Vanderbilt University,
Box 230 230 Appleton Place, Nashville, TN 37203-5721, USA
e-mail: doug.clark@vanderbilt.edu

G. Biswas
Department of EECS/ISIS, Vanderbilt University, Box 351824,Sta B, Nashville,
TN 37203, USA
e-mail: gautam.biswas@vanderbilt.edu

R.M. Luecht
Educational Research Methodology Department, University of North Carolina
at Greensboro, 240 SOE Building, PO Box 26179, Greensboro, NC 27402-6170, USA
e-mail: rmluecht@uncg.edu

P. Sengupta
Mind, Matter & Media Lab, Peabody College, Vanderbilt University, Box 230,
230 Appleton Place, Nashville, TN 37203-5721, USA
e-mail: pratim.sengupta@vanderbilt.edu

D. Ifenthaler et al. (eds.), *Assessment in Game-Based Learning: Foundations, Innovations, and Perspectives*, DOI 10.1007/978-1-4614-3546-4_10,

1996, 2012), and the broader twenty-first century skills recognized as critical for all citizens (NRC, 2010).

Digital games provide an ideal opportunity to support this richer view of science learning (Clark, Nelson, Sengupta, & D'Angelo, 2009; Clark, Nelson, Martinez-Garza & D'Angelo, submitted; Federation of American Scientists, 2006; Honey & Hilton, 2010). This chapter presents a model for operationalizing, supporting, and assessing students' progress and proficiency in alignment with these science proficiency goals. The focus of our approach is on prediction and explanation in game play by integrating computer-adaptive testing (CAT) technologies and hidden Markov modeling techniques to track students' activity and construct models of students' learning within a single-player game (although the approach can be extended to multiplayer games).

10.2 Background and Challenges: Games to Support Science Learning

The idea that games might provide affordances for science learning and inquiry is not idiosyncratic. In 2006, the Federation of American Scientists issued a widely publicized report stating their belief that games offer a powerful new tool to support education (Federation of American Scientists, 2006). The FAS report encourages governmental and private organizational support for expanded research into the application of complex gaming environments for learning. In 2009, a special issue of *Science* (Hines, Jasny, & Merris, 2009) echoed and expanded this call. Many studies have provided evidence for the potential of digital games to support science proficiency in terms of conceptual understanding and process skills to operate on that understanding (e.g., Annetta, Minogue, Holmes, & Cheng, 2009; Barab, Zuiker, et al., 2007; Clark, Nelson, Sengupta, et al., 2009; Coller & Scott, 2009; Dieterle, 2009; Hickey, Ingram-Goble, & Jameson, 2009; Holbert, 2009; Kafai, Quintero, & Feldon, 2010; Ketelhut, Dede, Clarke, & Nelson, 2006; Klopfer, Scheintaub, Huang, Wendal, & Roque, 2009; Moreno & Mayer, 2000, 2004; Nelson, 2007; Nelson, Ketelhut, Clarke, Bowman, & Dede, 2005; Steinkuehler & Duncan, 2008). Studies also show that games can support: (1) students' epistemological understanding of nature and the development of science knowledge (e.g., Barab, Sadler, Heiselt, Hickey, & Zuiker, 2007; Clarke & Dede, 2005; Neulight, Kafai, Kao, Foley, & Galas, 2007; Squire & Jan, 2007; Squire & Klopfer, 2007), and (2) students' attitudes, identity, and habits of mind in terms of their willingness to engage and participate productively in scientific practices and discourse (e.g., Anderson & Barnett, 2011; Annetta et al., 2009; Barab, Arici, & Jackson, 2005; Barab et al., 2009; Dede & Ketelhut, 2003; Galas, 2006; McQuiggan, Rowe, & Lester, 2008). There are, however, challenges involved in using games to support science learning in terms of assessment and in terms of helping players connect intuitive understandings developed through game play with explicit formal understandings. This chapter proposes an explanation dialog model to address these two challenges.

Challenge I: Assessment. One central challenge for game-based learning involves assessment. Specifically, pre–post multiple-choice tests, while exceptionally common,

Table 10.1 Challenges with standard pre-post approaches to assessment of learning in games

Challenge	Description
Standard pre-post tests cannot track learning processes within a game or activity	While they may, in fact, provide evidence of student learning, standard pre-post tests cannot provide critical information about the conceptual change processes involved (e.g., how students' intuitive concepts guided their answers and their play, what levels of scaffolding were most helpful, and how their emergent understanding guided their game play)
Standard pre-post tests require a large number of items to reliably assess a student's understanding	The span of items administered in the form of decontextualized format of summative assessment often results in test fatigue and disinterest by students, which results in added noise for which most statistical models do not account
Standard pre-post tests are costly in terms of time and opportunity because they are summative rather than purposefully or effectively instructional	Teachers are often not interested in allocating instructional time to the pretest, and if the curriculum/game is short enough, may not be interested in allocating instructional time to an extended post test that doesn't cover an extended span of their curriculum
Standard pre-post tests typically cannot assess extended problem solving	While the new science proficiency standards focus on students' ability to engage in deep extended problem solving involving the application of conceptual knowledge, most pre-post tests do not support or track extended problem solving
Standard pre-post tests often do not capture the connections between intuitive understanding and explicit formal understanding	Most multiple-choice tests focus only on explicit (and rote) representations of ideas. Tests of conceptual physics, such as the FCI, may focus on tacit understanding in their efforts to avoid assessing rote information, and in the process, may not assess students' ability to connect tacit understanding with explicit formal understanding

have many shortcomings in the context of games, such as assessing the richer forms of understanding and performance in science learning that will occur during game play (Clark, Nelson, Sengupta, et al., 2009; Clark et al., submitted). First, pre- and post tests only measure understanding before and after an intervention (i.e., the game)—pre-post tests do not track the processes of knowledge construction within a game or activity. Second, standard pre-post tests require a large number of items to reliably assess a student's understanding. Third, standard pre-post tests are costly in terms of time and opportunity because they are summative rather than being purposefully or effectively formative and, therefore, supportive of the learning process. Fourth, standard pre-post tests typically cannot assess extended problem solving. Fifth, and finally, standard pre-post tests often do not capture the connections between the intuitive understanding that students gain by playing a game and the formal, generalized understanding that students need to develop to become effective problem solvers in the domain of study. Table 10.1 explores these assessment challenges in greater depth. At the same time, other approaches to assessment in games for learning clearly need to be explored (Quellmalz & Pellegrino, 2009).

Challenge II: Connecting intuitive and explicit formal understanding. A second challenge area in games for science learning involves helping students connect the intuitive understandings they develop through game play with the explicit formal representations and concepts of the targeted science disciplines. Research on *Supercharged* (a 3D game in which players utilize and explore the properties of charged particles and field lines to navigate their ship through space), for example, found that students made significant learning gains on the physics post test, but only when the teacher collaborating in the research created activity structures outside of the game to engage students in predicting and explaining what was happening in the game and reflecting on connections of the tacit intuitive knowledge that the students were building through game play to the representations and concepts of the formal discipline (Squire, Barnett, Grant, & Higginbotham, 2004). Masson, Bub, and Lalonde (2011) showed similar outcomes where students appeared to develop intuitive understanding of aspects of the physics involved through the core game-play, but this intuitive understanding did not help students on subsequent assessments that tested explicit formal understanding. Work on SURGE (another conceptually integrated game where students use physics principles to navigate through space to achieve a variety of goals linked to a rescue theme) focused on integrating supports for connections between intuitive understanding and explicit formal physics representations and concepts showed significant gains on test items based on the Force Concept Inventory (FCI), a prominent conceptual test of undergraduate physics understanding about force and motion. Studies with SURGE also showed that further scaffolding is needed to help students build stronger connections between the intuitive understanding developed through game play and the targeted explicit formal concepts. (Clark et al., 2011). Thus the rich integrated conceptual understanding and ability to explain and apply that understanding targeted by the new science proficiency standards requires deeper learning behavior analysis and translating this information into appropriate metacognitive scaffolding within games for learning.

10.3 Framing a Solution to the Challenges

We frame our solution, the explanation dialog model, through two sets of cognitive goals as outlined in this section.

10.3.1 Cognitive Goal 1: Leverage Explanation Within Games to Support Learning and Assessment

If our goals for learning and assessment move beyond transfer and recall of rote information to include the ideas about science proficiency, we need to engage the player actively in processes of thinking that parallel the new science proficiency goals. We propose that engaging students in explanation related to problem solving offers excellent leverage for both the learning and assessment.

Research on self-explanation by Chi and others provides clarity into the value of explanation for learning (e.g., Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi & VanLehn, 1991; Roy & Chi, 2005). A recent review of research on students' self-explanation reports that self-explanation results in average learning gains of 22% for learning from text, 44% for learning from diagrams, and 20% in learning from multimedia presentations (Roy & Chi, 2005). Encouragingly, research by Bielaczyc, Pirolli, and Brown (1995) shows that instruction that stresses explanation generation improves performance even after the prompts to explain are discontinued. Mayer and Johnson (2010) have conducted preliminary work in embedding self-explanation in a game-like environment with encouraging results that include gains on transfer tasks.

This emphasis on explanation is mirrored in research on science education. Work by White and Frederiksen (1998, 2000), for example, demonstrated the value of asking students to reflect on their learning during inquiry with physics simulations. This emphasis on explanation is often accompanied with prediction (e.g., Grant, Johnson, & Sanders, 1990; Mazur, 1996 reviewed more generally in Scott, Asoko, & Driver, 1991), promoting metacognition, learning, and reflection (e.g., Champagne, Klopfer, & Gunstone, 1982), enabling conceptual change (Borges, Tecnico, & Gilbert, 1998; Kearney, 2004; Kearney & Treagust, 2000; Liew & Treagust, 1998; Palmer, 1995; Shepardson, Moje, & Kennard-McClelland, 1994; Tao & Gunstone, 1999), while also providing a useful tool for probing and diagnosing students' conceptions of science facts and monitoring conceptual change (Liew & Treagust, 1995, 1998; Searle & Gunstone, 1990; White & Gunstone, 1992).

A growing body of research and scholarship on games and cognition emphasizes cycles of prediction, explanation, and refinement as the core of game-play processes (Games-to-Teach Team, 2003; Salen & Zimmerman, 2003; Wright, 2006). Few games provide coherent structures for externalizing and reflecting on game-play; more often, such articulation and reflection occur outside the game, through discussion among players and participation in online forums (Gee, 2003/2007, 2007; Squire, 2005; Steinkuehler & Duncan, 2008). We propose that supports for this kind of articulation and reflection can be integrated within the game itself.

10.3.2 Cognitive Goal 2: Constraint-Based Thinking Versus Model-Based Thinking

One of the purposes for integrating explanation into a game is to catalyze model-based thinking. Parnafes and diSessa (2004) explored players' thinking in a game-like simulation called *NumberSpeed*. Their research showed that players sometimes engaged in thinking very locally through simple processes of covariation (constraint-based reasoning), and at other times, engaged in deeper thinking about the underlying relationships and components to make more principled or model-based accounts and solutions for the challenge (model-based reasoning). They defined constraint-based reasoning as "using a set of heuristics to meet the problem constraints, usually using simple covariation" (p. 265). Constraint-based thinking involves means-ends strategies focusing on local comparisons and matching, simple motion principles, or pure covariation focusing on a small number of the problem constraints or parameters. Model-based reasoning, as Parnafes and diSessa explain, involves “creating a mental model of the whole scenario of motion, and mentally running the model to reason about the motion situation” (p. 268) to examine plans and modify or develop alternative plans in pursuit of an integrated qualitative solution based on the model.

While constraint-based thinking is fine in so far as it supports the development of model-based thinking, model-based thinking is ultimately needed for deep and integrated understanding. This makes sense from an elemental perspective on conceptual change (e.g., Clark, 2006; Clark, D’Angelo, & Schleigh, 2011; Clark & Linn, 2003; diSessa, 1993, 1996; Hammer, Elby, Scherr, & Redish, 2005; Hunt & Minstrell, 1994; Minstrell, 1982, 1989; Minstrell & Kraus, 2005; Sengupta, 2011; Sengupta & Wilensky, 2009, 2011). According to these perspectives, learning occurs as people sort through and refine their ideas as they build and refine connections between the ideas. If the games only demand constraint-based reasoning of the player, very little substantial reorganization and revision of the player’s ideas is required in comparison to games that require model-based thinking. Similarly, from an assessment perspective, if games only elicit constraint-based thinking, we cannot assess what we care most about: students’ ability to connect intuitive and explicit formal understandings in a principled manner to solve problems.

10.4 High-Level Model: Integrating and Assessing Explanation in Game Dialog

In addition to the cognitive goals for the explanation dialog outlined above, there are also driving goals from a game design perspective. We cannot just have students write predictions and explanations in a journal, for example, because that would destroy the flow of the game experience. Our intention is to fit explanation generation into the game narrative, in a way that preserves narrative space (Salen & Zimmerman, 2003), allows for identity construction and agency (Gee, 2004; Pelletier, 2008), and respects learners’ expectations and aims regarding the essence of play (Caillois, 1961; Huizinga, 1980). All of these are important elements of games and play, which, it is hypothesized, can be disrupted by assessment (Shute, Rieber, & Van Eck, 2011). We propose that explanation generation can be integrated into the dialog of a game by encouraging self-explanation in the dialog between the players and the characters in the game. What might this look like? We outline a general model on how explanation might be enacted in game dialog.

For the purposes of our model, we will assume that there is a “core” game around which a set of “explanation” games will be developed. The core game focuses on a science topic, and the game is structured in a way that the player has to apply science concepts to navigate or work toward the established goals for the game. There are many commercial and educational games types that could provide the basis for a core game (see Fig. 10.1). Many of these are physics games (e.g., Angry Birds, Crayon Physics, Supercharged, SURGE, Switchball, Gravitee), but good examples also exist in chemistry (e.g., SpaceChem) and biology (e.g., SimAnt,

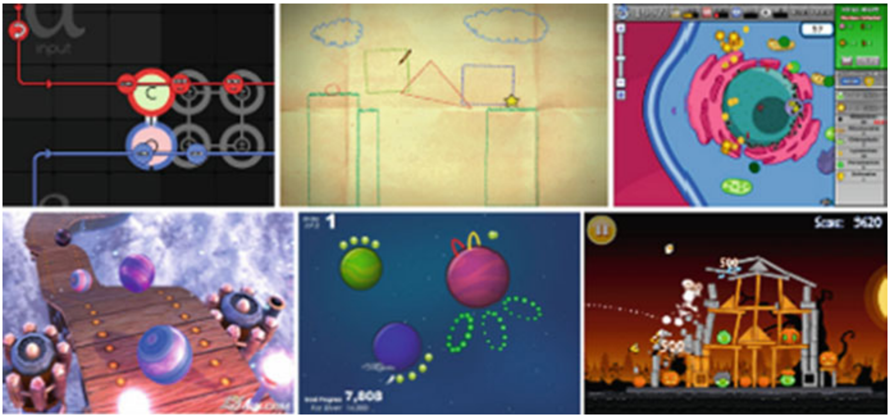

Fig. 10.1 Many recreational and educational games could provide the basis for the core game by engaging the player in applying science concepts to navigate or work toward an established goal. Examples (from *left* to *right*, *top* to *bottom*) include *SpaceChem*, *Crayon Physics*, *CellCraft*, *Switchball*, *Gravitee*, and *Angry Birds*

SimLife, CellCraft). Pedagogical agents and scaffolding might be layered on top of these mechanics to support players in identifying relevant and important ideas in the game play of the core game. One of the primary goals of the core game is to facilitate developing an intuitive understanding of the science concepts through game play (similar to learning by doing). Games, by creating engagement and flow (Csikszentmihalyi, 1991), have traditionally done well in this regard (Clark, Nelson, Sengupta, et al., 2009; Clark et al., submitted).

We then create a parallel explanation game to help students formalize and generalize their models so that they can be applied to a wide range of related problems (i.e., problems that are based on the same set of primary science principles). Levels of the explanation game are interwoven between levels of the core game. We believe that these explanation game levels (or "challenges") can support assessment and also build connections between the intuitive understandings the students develop through playing the core game and explicit targeted formal concepts and representations of the discipline. Essentially, the player first plays the core game "for themselves." Their play in the core game is scaffolded with prompts and suggestions from mentor agents. The player then takes on the role of a mentor in the explanation game. In the explanation game, the player teaches or helps one or more computer-controlled nonplayer characters (NPCs) to solve specific targeted challenges in the game environment. At its core, the explanation game:

1. Engages the student in identifying solutions to specific challenges that highlight and explore one or more core conceptual components from the science domain that were targeted in the core game.
2. Engages the student in developing and explaining this core game solution in a more general form, and at multiple levels of abstraction, with the goal of supporting the student in making the connections between the intuitive ideas developed during the core game play and the explicit formal versions of those ideas.

In this explanation game, players are asked to craft effective explanations that highlight and clarify the formal science ideas (e.g., Newton's Laws and associated key ideas of kinematics) to aid the cause of these characters and thus earn the player additional recognition and in-game rewards. To structure the explanation dynamic in a meaningful, appealing, and engaging way, players have the opportunity to explain and justify their strategies and the concepts underlying those strategies to NPCs in order to (a) convince the NPCs to adopt these solutions and (b) help the NPCs successfully overcome similar focused challenges. The challenges faced by the characters will often be presented as contrasting cases tied to common misconceptions (Bransford & Schwartz, 1999; Schwartz & Martin, 2004). Students can get immediate feedback on the quality and correctness of their explanations by observing how well their characters perform when they use the knowledge implied by the explanation to solve their assigned mission tasks. In previous work (e.g., Biswas, Leelawong, Schwartz, Vye, & The Teachable Agents Group at Vanderbilt, 2005; Schwartz, Blair, Biswas, & Leelawong, 2007), we have found that, if properly scaffolded, this motivates the student and helps to direct their attention to mastery as opposed to performance goals (Pintrich, 2000).

Therefore, the explanation game is a manifestation of the hypothesis that, by providing players with multiple meaningful opportunities for explanation embedded within the game, as well as appropriate tools and scaffolding for developing and assessing these explanations, the game experience will foster deep learning of complex curricular science concepts. This explanation activity is framed in terms of a dialog with characters in the game environment. This game design element is familiar to players and also an efficient way to pace and structure the natural flow of information. To make the explanation tasks meaningful and to embed them within the game narrative, the explanation opportunities will be couched in terms of aiding other characters in solving similar puzzles as the ones the student has just solved in the core game. In many ways this taps into the learning by teaching paradigm (Bargh & Schul, 1980; Biswas, Schwartz, Bransford, & The Teachable Agents Group at Vanderbilt (TAG-V), 2001) as well as the self-explanation paradigm (e.g., Chi et al., 1989; Chi & VanLehn, 1991; Roy & Chi, 2005).

In addition to the literature reviewed earlier, we are building on design paradigms focused on adding an explanation task following a feedback event (Mayer & Johnson, 2010). While playing an electronics quiz-based environment that Mayer and Johnson defined as being game-like, students were tasked with answering explanatory questions posed as circuit diagrams. We believe that the results of this study suggest that asking students to perform some activity that connects the general scientific principles with the task the student just performed (whether they were successful or not) can be very conducive for deep model-based learning in a true game context, particularly if the explanation is integrated within the fabric of the game and, crucially, if the scaffolding surrounding the task is more responsive to students' thinking. That is, we believe that it is through designing scaffolds for supporting self-explanations throughout the game that we can foster model-based reasoning (Parnafes & diSessa, 2004) in students.

We have discussed how the model above might support learning in terms of the discussion of our cognitive goals, but how might we leverage CAT and hidden Markov models (HMMs) to analyze students' explanations and game-play data in real-time to model understanding? More specifically, how could a game developed using this model track and assess a "just-in-time" view of students' understanding within the navigation and explanation components of the game? We propose that game-play data in both components of the game, generated by learners as they play the game, could provide an avenue of assessment that would allow valid inferences about learning behaviors and strategies, and therefore, provide a richer interpretation of learning outcomes (Shute et al., 2011). The goal of the model would be to support formative and summative assessment within the game in terms of the player's understanding of the target formal physics ideas early in the game, how and when that understanding evolves, and the degree of formal understanding the player has developed by the end of the game. These assessment models could then provide diagnostic information to (a) support just-in-time adaptive scaffolding in terms of the actions and suggestions that NPCs in the game make as pedagogical agents to support player learning during the game as well as the order and nature of the levels that the player encounters and (b) provide diagnostic information to researchers and teachers to support inferences about learning and to guide subsequent instruction.

10.5 Structure of the Explanation Game

We now outline a possible structure for the explanation game to explore whether it might be possible to integrate explanation into the dialog in an engaging game-like manner. This example will also facilitate discussion in subsequent sections about the integration of CAT and hidden Markov modeling techniques to assess progress and understanding in real time.

Essentially, a level (or "challenge") in the explanation game involves a multitiered challenge that spans a few minutes (maybe 1 min for a player who has a firmer grasp of the concepts underlying the challenge and 3 min for a player who has a less firm grasp of those ideas). The challenge is selected to be (a) a specific difficulty that is adjusted based on the player's previous performance and (b) a challenge the player has not yet encountered.

Each challenge consists of a sequence of roughly four tiers that engage the player in identifying or proposing a solution to a dilemma faced by the computer-controlled NPCs and then justifying or explaining that answer to the NPCs at multiple levels of conceptual abstraction. Advancing in the challenge requires learners to think about physics concepts in a general format using representations that are general and create clear contrasts between different problem types. For example, a challenge might open with one of the NPCs framing an emergency scenario that the player has to solve (e.g., their ship is about to crash). This scenario could be portrayed using the engine from the core game to show what the NPCs are doing in the game and the potential impending problems they face in solving the problem. The player then proposes or identifies a solution and justifies this choice to the NPCs to convince them to adopt the solution. The game engine then models the solution. The player is rewarded appropriately based on the efficacy of the solution for the NPCs' dilemma.

At each tier of the challenge, if the player creates or selects a nonproductive path or strategy, the player receives feedback and support in revising their choices. Thus all players are scaffolded in creating and justifying a functional solution. Scoring is, therefore, based on how efficiently a player moves through the challenge rather than whether or not the player reaches a productive solution (because all players are scaffolded in eventually achieving a productive solution). Each challenge is conceptualized as a learning opportunity that provides formative feedback, rather than simply assessing whether or not a player can solve the challenge.

From a programming perspective, the simplest version of a challenge might employ the standard computer game conventions for dialog. In the standard convention, when a player is asked to respond to an NPC's question in this standard convention, the player is presented a list of choices that are typically text only. This version of explanation is most closely related to the approach pursued by Mayer and Johnson. We also envision more visual and flexible approaches for player input, using combinations of text, images, and diagrams, along with other modes of framing a solution or explanation, either based on causal concept maps (e.g., Leelawong & Biswas, 2008) or discipline-specific representations such as free-body diagrams, flowcharts, or circuit diagrams. Thus, the player then goes through the tiers of the challenge, selecting the appropriate recommendations to help the NPC resolve their perilous situation. When a player chooses or designs a productive approach to a tier, the player moves to the next tier to build and extend on their initial explanation. When players choose a less productive approach to a puzzle, the dialog continues and they get feedback to help them understand the implications of their initial choice and a chance to make a different choice. Thus, a player will ultimately work his or her way through the productive approaches at each puzzle, but may require more or less feedback to do so. This might involve creating a feature in the game where students can unwind (go back a number of steps) and rethink their solution or explanation in light of problems they ran into with their previous solution or explanation.

A player's score in the game is a function of various factors including the number of steps and choices that they use to solve the challenge. This is represented as a counter (or "clock") advancing one step with each choice players make, with extra "time" allotted for achieving a more principled solution to a problem. If the player pursues too many nonproductive approaches, the counter will increase beyond a certain threshold and the game will inform the player that, for instance, their advice will reach the NPCs too late. The player then receives a reward (e.g., a medal) for the challenge depending on their score. A bronze medal, for example, might be awarded for getting to the end with much help and multiple missteps. A silver medal might be awarded for making few missteps in finding and explaining a solution, and gold medal might be awarded when the player can identify and explain the solution perfectly. The flow of the challenges is dynamic; if the player earns a bronze medal, for example, the difficulty of their next challenge might be adjusted slightly downward, a silver medal might keep challenge difficulty roughly the same, and a gold medal might result in adjusting the difficulty of the next challenge slightly higher. To maintain the pace of the game experience, a challenge should take between 1 and 3 min, depending on how much scaffolding the player requires.

Ideally, the game would contain a large library of challenges that have been validated and tested for item difficulty, as well as how they load onto any subscales of

Table 10.2 Possible design principles for structuring an explanation game that functions as an engaging game-like experience, a scaffold for deep learning, and the basis for assessment

1. Each time a player attempts a challenge is considered a trial. A trial involves one session of a player trying to navigate through the dialog forks of one challenge to reach a solution and supporting explanations that will work for that challenge
2. At the beginning of each trial, the game queries the previous trial data for that student to compute the difficulty of the challenge to present and cross-indexes with the catalog of challenges and their difficulties
3. At each fork of a given challenge problem, there is a prompt presented. The player then chooses or specifies a response. These responses can be in open form or presented as a closed set of choices. The response drives which branch of the fork the player moves down
4. An incorrect choice will result in returning to an earlier tier to choose again, with no additional penalty. When a player is returned to a fork where they made an incorrect choice, the incorrect choices are highlighted in some way or are not displayed at all
5. The game engine should be flexible enough to allow other interfaces to be added at given forks to allow the player to create/choose an alternative response (e.g., a concept map or free-body diagram). The general goal is to develop interfaces that have simple enough combinatorial complexity that logical and mathematical operators can be used to map ranges of answers to branches of a fork
   a) For a concept map format, the configuration possibilities for a given tier should be small enough that the combinations could be mapped onto the branches of the fork using logical operators
   b) The free-body diagram interface could use a combination of logical operators as well as computational algorithms, and the resultant composite vector could be mapped onto other branches for appropriate conceptual or procedural feedback
   c) Students could choose sentence fragments from a series of pull-down menus to create an explanation that the software could then assess and act on using a script with formal and mathematical operators and computations that operated on the student's choices to create a score for that explanation (Clark, 2004; Clark, Nelson, D'Angelo, & Menekse, 2009; Clark & Sampson, 2007)
6. The tiers of each challenge should be designed to guide students in crafting principled explanations. A principled explanation is defined as making a choice and providing both abstract reasons in terms of general principles, and specific values for the variables in play
7. We suggest that proper design consideration be given to authoring tools for the creation of challenges, including the use of templates. Ideally, the author can specify the layout for each tier of the challenge in its data file, and the kind of interfaces that may be included, in a straightforward manner. This might involve a separate data file for each challenge and a catalog file that the game engine can reference to identify appropriate challenges in terms of difficulty and focus

interest. A broad range of item difficulties and subscale profiles is a key feature to enable the CAT module. Additionally, several kinds of data would be recorded for each challenge undertaken by a player, so that both assessment and model-based feedback functionalities (see later sections) can be performed. The information stored should be extensive enough to reproduce as much of the player's actions as possible, including not only success/failure and number of trials, but also total time spent on each challenge, time spent on each branch of the challenge. For those interested in additional details, Table 10.2 outlines possible design principles for structuring an explanation game that functions as an engaging game-like experience, a scaffold for deep learning, and the basis for assessment.

10.6 Computer-Adaptive Testing Techniques to Drive Dialog and Analysis

This approach to supporting the development and use of learners' explanations within dialog lends itself to designing students' interactions and tracking students' progress through computerized adaptive testing techniques. CAT is being increasingly used in educational assessment setting to improve measurement efficiency and accuracy. There are numerous examples of successful, large-scale CAT programs, such as the ACCUPLACER postsecondary placement exams (operated by the College Board), the Graduate Record Exam (Eignor, Way, Stocking, & Steffen, 1993), and the Armed Service Vocational Aptitude Battery (Sands, Waters, & McBride, 1997). As the explanation game progresses, a CAT algorithm could sequentially select and administer challenges matched in difficulty to that student's apparent level of understanding and explanatory skills. Essentially, challenges are the "items" that the CAT functionality administers. For example, as the student performs better along a particular dimension of measurement interest in the game, more difficult challenges are presented. Conversely, worsening performance will cause easier challenges to be administered.

In CAT, the difficulty of every item is directly considered in scoring, since most CATs are based on item response theory (Lord, 1980). IRT has a long history of use in educational settings (e.g., Yen & Fitzpatrick, 2006), especially for scaling and equating end-of-course and end-of-grade tests used by most states. IRT relies on a probabilistic model that related the responses to an underlying proficiency scale, typically referred to as θ (or, the Greek "theta"). A commonly used IRT model is the three-parameter logistic model, $\text{Prob}(u_i = 1 | \theta; a_i, b_i) \equiv P_i(\theta) = \{1 + \exp[-a_i(\theta - b_i)]\}^{-1}$. This function generates a probability curve denoting the likelihood that an examinee having a proficiency score, θ, will correctly answer item i, which has a sensitivity or discrimination parameter, a_i, a difficulty parameter, b_i, and a lower asymptote parameter, c_i, where the latter is often conceptually assumed to be related to guessing behaviors on multiple-choice test items by lower proficiency examinees. Different items have different a_i, b_i, and c_i parameters, and these differences are taken into account in scoring. When all of the items are calibrated to a common scale—conceptually similar to calibrating weights or laboratory equipment—we can estimate examinees' proficiency scores, even when they take tests that differ in difficulty. For example, Fig. 10.2 shows the expected number-correct scores for three 25-item tests: an easier test, a moderate difficulty test, and a difficulty (hard) test. A number-correct score of 15 (60% correct) on the easy test maps to a θ score of slightly less than −1.0 (the actual estimate is approximately −1.15). The same number-correct score produces a θ score of about +0.35 on the moderate test and a θ score of approximately +1.15 on the most difficult test. In other words, the calibrated item statistics automatically adjusts the scoring for the difficulty (and other characteristics) of the test.

Calibrated IRT item statistics are used in CAT to actually target the difficulty of the test form to the apparent proficiency of the examinee. In principle, a CAT test delivery system could produce a unique test form for every examinee. That would

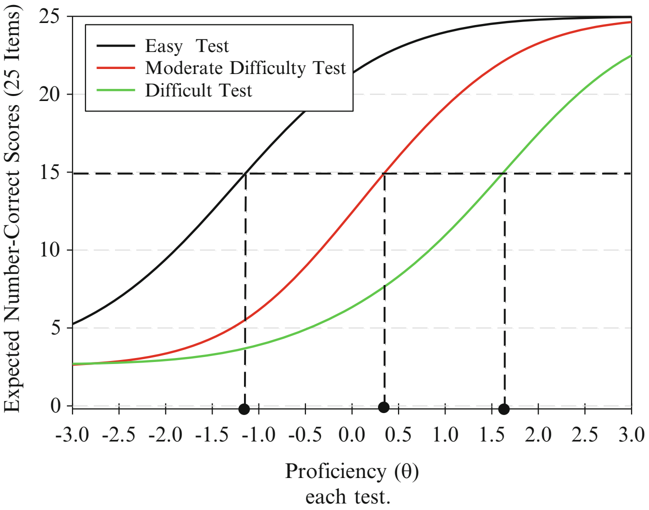

Fig. 10.2 Expected scores for three 25-item tests: easy, moderate, and difficult. Proficiency scores are mapped corresponding to a number-correct score of 15 on each test

imply a unique expected test score curve, similar to the unique expected performance curves in Fig. 10.2, for each examinee. An examinee's actual performance on his or her CAT automatically takes the item difficulty into account. Therefore, getting easier or more difficult items does not penalize the students. Rather, each new, adaptively selected item actually improves the precision of the estimated score profile by systematically reducing measurement errors.

Figure 10.3 shows the proficiency scores for three examinees: one examinee with relatively low proficiency, one with medium proficiency, and one with high proficiency. The item difficulties are shown by a "*" and track fairly closely with the estimated proficiency scores. This is the CAT targeting the item selection to each examinee's proficiency score. The CAT starts near the center for each examinee and then diverges toward the examinee's apparent proficiency. Also, the error bands around each score point continue to shrink in size, demonstrating greater confidence in the accuracy of the score estimates as more items are administered.

Unlike a conventional fixed test form, where every examinee sees the same items, an individually tailored CAT is usually far more precise and takes less testing time than a conventional test form (van der Linden & Glas, 2010). Through enhancements to the CAT item selection algorithm and scoring process involving multidimensional item response theory models (e.g., Luecht, 1996; Segall, 1996; van der Linden & Glas, 2010), it is entirely possible to develop highly informative multidimensional profiles of proficiency that truly are formative and diagnostic in nature. This need for a multidimensional perspective of strengths and weaknesses in formative assessment settings has only recently been demonstrated (e.g., Leighton & Gierl, 2007). Here, that multidimensional perspective can be efficiently applied to simultaneously measuring multiple learning progressions of complex constructs.

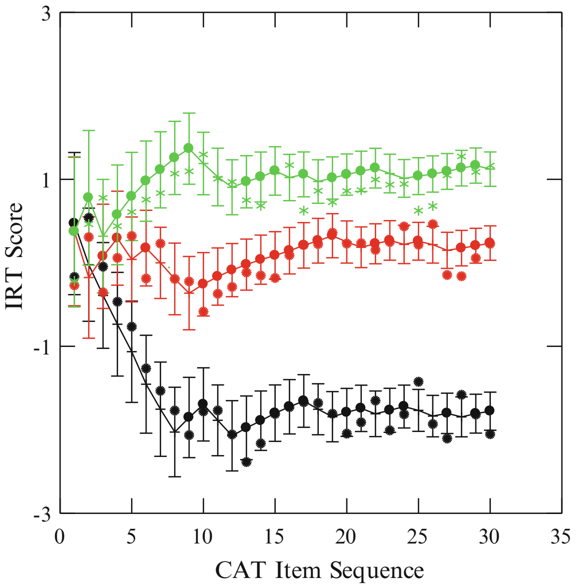

Fig. 10.3 Patterns of estimated θ Scores for three examinees (low proficiency, medium proficiency, and high proficiency) showing decreasing errors of estimate across the 30-item CAT sequence (*closed cirle* denotes a correct response; – indicates an incorrect response)

Essentially CAT technology could adjust the difficulty of the challenges that a player encounters within the explanation game. This would allow the game engine to build an ongoing evolving model of the player's current understanding via a multidimensional knowledge and skill profile. As described in the preceding section, challenges within the game would ask the player to explain to other characters in the game why something is happening, why something isn't working, or how to solve a problem systematically using the underlying concepts. These challenges could be developed as general templates so that multiple variants of the challenge could be generated. Variants could be as simple as switching numbers in the question, such as the mass or initial speed of an object, or could involve other variants, such as direction or combination of forces involved. Including multiple variants would allow a student to receive the same challenge (essentially) at a later time as a different variant if they do not answer it correctly in the first encounter and to prevent the sharing of answers in a classroom environment. Thus, integrating this CAT structure to select challenges in the explanation game would allow tracking and rechecking for evolving progress and understanding across game play.

Challenges will be developed, piloted, and ultimately calibrated to each of those multidimensional scales using item response theory, as noted above. Validity of the challenges will also be tied to the cognitively oriented construct maps associated with each of the scales to help ensure that proper interpretations of performance can be made—that is, inferences based on the specific explanations that the players provide. Once an item pool of IRT-calibrated challenges is developed, the explanation software could adaptively select and administer challenges during the game, with provisional scores helping the CAT algorithm to make the best challenge choices insofar as maximizing the precision of a multidimensional proficiency profile (scores and explanations of performance) generated for each player.

10.7 Hidden Markov Modeling to Track Game Play

CAT techniques can track and direct dialog within the explanation game, but another approach is required to track students' learning behaviors and the strategies they employ to develop understanding and problem solutions in the core and explanation games. Individual game play is a reflection of a complex interplay of behavioral, cognitive, and socio-constructivist elements that resist a simple causal construct. This resonates with the theoretical frameworks for game-based learning proposed by Amory (2006), Gunter, Kenny, and Vick (2008), Squire et al. (2004), and others. Hidden Markov modeling techniques provide strong affordances for capturing patterns in this learning as state-based models (Li & Biswas, 2002; Rabiner, 1989). Technically, a HMM describes a probabilistic state machine that describes a phenomena or behavior that evolves over time. The behavior is modeled in a compact form as a set of finite discrete, hidden (not directly observed) states, and probabilistic transitions between these states. The manifestation (or observation) of this behavior are observed symbols or numbers that are defined as the output corresponding to these states. HMMs have been successfully used in speech synthesis and recognition, gesture recognition, and for analyzing protein sequences in bioinformatics (Brand, Oliver, & Pentland, 1997; Juang & Rabiner, 1991; Krogh, Brown, Mian, Sjolander, & Haussler, 1994).

Figure 10.4 shows a hypothetical student's learning behaviors represented as a three-state HMM. While the three states cannot be directly observed, they can be inferred from the students' activity sequences. We can then examine the probabilities of producing each action in a state in order to interpret the meaning of that state. For example, the information-gathering state derives its name and meaning from the activities produced in that state (i.e., the state's output), such as reading resources and taking notes. Similarly, the map building state is associated with activities that include adding, deleting, and modifying concepts and links to create a concept map representation for the topic of study. The monitoring state is defined by actions like asking questions to see if one has understood a concept and taking quizzes provided by an instructor to check how one's performance is on a given topic of study. The transitions in the example model indicate likely sequences of actions a student may take. For example, a student will likely perform a map building action after an information gathering action with a probability of 0.3. But the student may continue with the information-gathering task (with a probability of 0.5), and less likely, a monitoring action to check if their map is correct (with a probability of 0.2).

We have used the HMMs to derive concise representations of student learning strategies and behaviors (Biswas, Jeong, Kinnebrew, Sulcer, & Roscoe, 2010; Jeong & Biswas, 2008). Algorithms for learning an HMM from output sequences are

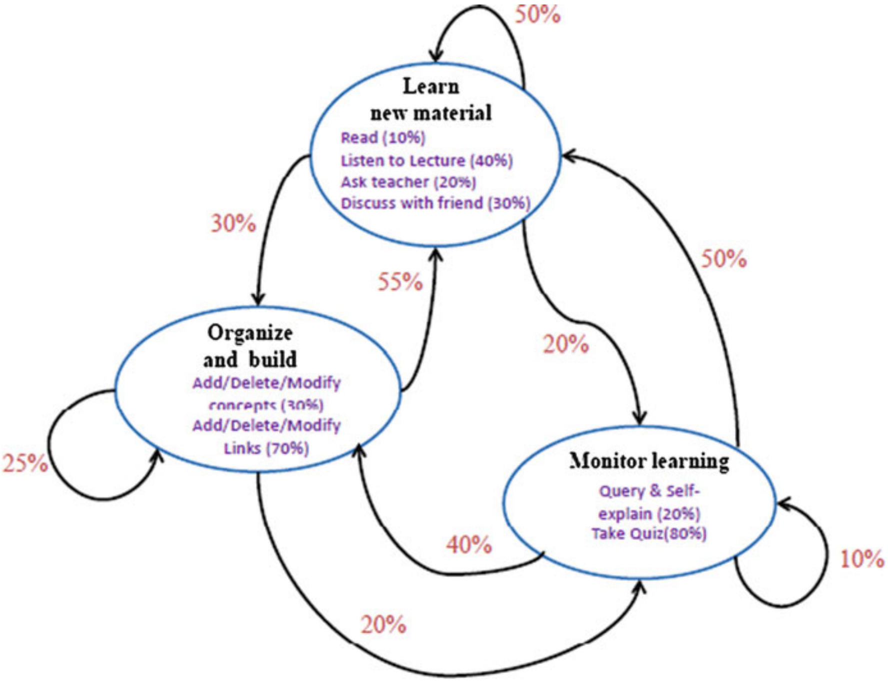

Fig. 10.4 An example of HMM structure describing students' learning behaviors

well-known, but require appropriate configuration/initialization parameters for effective use (Rabiner, 1989). Specifically, HMM learning algorithms require an initial HMM whose parameters are then modified to maximize the likelihood of producing observed output sequences. In particular, the number of states in the HMM and their initial output probabilities are important parameters that influence the structure and interpretation of the learned HMM.

Essentially, HMMs can be learned from observing students' activity sequences during core game play as well as the moves they make in the explanation game. The derived HMMs provide a probabilistic model of students' behavior patterns and their related learning outcomes. The model is derived from sequences of observed activities and the consequent results they produce in the game environment. Such models provide us with a framework for characterizing good vs. suboptimal strategies that students employ in learning and problem solving. A suboptimal strategy may be reflected as a trial and error or a guessing approach to obtaining a problem solution, whereas a good strategy may manifest as deriving a problem solution step by step, checking whether a science principle or concept has been correctly applied for each step and reflecting on intermediate results to check if the expected solution is being generated.

Previously, we have used HMMs to model student learning strategies and their interrelationships in an environment where students learn by teaching a computer agent (Biswas et al., 2010; Jeong & Biswas, 2008). In this work, HMMs have successfully captured underlying structure in sequential data (students' activity sequences and learning outcomes as computed by CAT items) as hidden states with probabilities of producing observable outputs (actions/outcomes), as well as probabilities of transitioning between those states. States represent higher-level cognitive states, such as informed editing of concept maps and using explanation structures to probe the correctness of a causal path structure (for details, see Biswas et al., 2010). Therefore, HMMs learned from student activity sequences can provide an overview of common behaviors/strategies employed by individual students or a set of students during learning, as well as the likelihood of transitioning between strategies while they are working on their learning, explaining, and problem-solving tasks in the game environment.

In more recent work, HMM algorithms have been employed as a bottom-up tool for learning students' behavior sequences from their log activity data (Biswas et al., 2010). This has been combined with top-down modeling approaches, informed by the metacognition and self-regulated learning (SRL) literature (e.g., Azevedo, 2009; Pintrich, 2000; Schraw, Crippen, & Hartley, 2006; Schwartz et al., 2009; Winne & Hadwin, 2008; Zimmerman, 2001) to map observed learning behaviors to (suboptimal and optimal) strategies that students employ for learning and problem solving. The ability to construct and use such models online provides a framework for developing explanatory dialog structures and feedback mechanisms that scaffold and support student learning during their game play activities. Not only can this provide critical diagnostic information to researchers and teachers, but this can also be used to create rich social interactions with the game characters to support conceptual development, metacognition, and engagement/immersion.

In the case of our model, a HMM trained on the moves and choices students make in the game will reveal the (hidden) aggregate state-based model that explains those choice and activity sequences. By monitoring the students' actions and performance associated with the states, the HMM can be used to make inferences, such as recognizing that students are employing "trial and error" methods to determine resultant forces, or a state where students systematically experiment with collisions to study conservation of momentum. Furthermore, since the HMM reports transition probabilities between states, these techniques can track how students combine the use of strategies to solve bigger problems. Longitudinal tracking of students across multiple problems could monitor how states and transitions change over time, providing the basis for inferences about changes in the underlying assumptions that guide students' thinking. Crucially, HMM analyses are performed using time-structured data (not just snapshots of data captured at set moments during the intervention). Hidden Markov modeling is thus well-suited to applications within games for learning, where the dynamic features of students' play and learning evolution can be captured as learners become proficient in the game and the operating principles behind it.

Crucial to all of this is determining the level of abstraction or detail for the input sequences from which the HMM is developed. This could range from individual keystrokes and mouse clicks a student makes (this may be too low-level for our purposes) to more aggregate representations, where small sequences of observed activities are considered to be related and represented as a single activity. HMMs can be developed from any such sequence of activities. The interpretation of the states generated will depend on the level of detail chosen for the activities that make up the sequences. Relevance measures or sequence-mining methods can be used, for example, to define and categorize the primitive actions on which the behavior analysis is based (Biswas et al., 2010; Kinnebrew, Loretz, & Biswas, in press, 2011). Depending on the level of detail chosen for the activity sequences, the hidden states of the HMM or state transitions can be interpreted as behaviors, which can be further mapped onto cognitive and metacognitive strategies that the students apply during their game playing and problem solving. Building more abstract sequence descriptions after the first round of interpretation using sequence mining can provide a framework for generating more aggregate behavior models.

Once interesting patterns and the HMMs have been developed or "learned" with some consistency, they can be used within the game environment to trigger various forms of scaffolding and feedback to support student learning. For example, detection of suboptimal strategies for learning may trigger a suggestion from a peer or mentor agent that guides the student to think of better strategies they may employ for learning and problem solving.

10.8 Connections Between Computer-Adaptive Test Scores and Hidden Markov Model

The combination of computer-adaptive test data and behavior analysis and interpretation can guide the level of feedback provided to the user, and subsequently guide students through learning trajectories (e.g., choice of topics and problems) that help them optimize their learning performance. For example, HMM learning could further analyze the impact of the computer-adaptive test assessments in combination with other game play data. The idea is to extract activity sequences that are linked to scale and subscale assessments and performance. A number of different methods may be applied. For example, one could employ clustering methods to group students by performance in the multidimensional space of computer-adaptive test scores. One could then extract activity sequences by group and derive HMMs for comparative analysis of the behavior structure for each group. The interpretation of the comparative behavior analysis along with the computer-adaptive test profiles by group could provide a rich framework for designing context-relevant argument and dialog structures for providing scaffolds and feedback to support and improve student learning. An alternate approach that could provide much finer-grained analyses of behaviors is sequence mining (Agrawal & Srikant, 1995) in a manner that focuses on fine-grained differential analyses of behavior patterns employed by groups of students (Kinnebrew et al., in press, 2011). These methods could then form the basis for tracking scales and subscales within the game environment.

10.9 Final Thoughts: Does the Proposed Explanation Dialog Model Address the Challenges Outlined for This Chapter?

The proposed explanation dialog model addresses the five concerns raised about standard pre-post approaches to assessment. First, the model tracks learning processes within the game rather than simply before and after. This allows insights into students' learning processes and provides opportunities for real-time scaffolding based on the formative assessment. Second, the explanation dialog does not require a large number of items at the beginning of the game because it integrates initial assessment into the first parts of the game and instead measures progress across the whole game rather than focusing only on static pictures of pre and post performance. The explanation dialog does not require large numbers of items at the end of the game to reliably assess a student's understanding because the embedded assessments have already created a detailed profile of the student's understanding that can be refined through a finite number of additional computer-adaptive summative questions. Third, this model is not costly in terms of time and opportunity because the assessment activities are purposefully formative and instructional, rather than solely summative. Fourth, the explanation dialog could assess extended problem solving rather than isolated decontextualized problems. Fifth, and finally, the explanation dialog provides an approach for capturing and promoting the connections between intuitive understanding and explicit formal understanding.

In terms of the challenge of helping students connect intuitive and explicit formal understandings, the explanation dialog structures the explanation game assessments entirely around specific challenges and common misconceptions. The goal of these challenges involves (a) engaging the student in identifying a solution that highlights a core science concept targeted by the core game and (b) engaging the student in explaining the solution at multiple levels of abstraction, with the goal of supporting the student in making the thinking and connections between the intuitive ideas from the core game and the explicit formal versions of those ideas from the discipline. In doing so, the explanation dialog builds on research on self-explanation and model-based thinking from the psychology, learning sciences, and science education literatures.

We thus claim that the proposed explanation dialog holds the potential to address the challenges to game-based learning outlined at the onset of this chapter. There are, however, limitations and trade-offs. The largest of the current limitations is the simple multiple-choice nature of the standard branching dialog tree (common to games), which does not provide a fully open-ended format. Essentially, writing multiple-choice items that test rote knowledge is simple (as evidenced by the vast proliferation of such items in typical multiple choice tests), but construction of deep conceptual dialog challenges using this format involves the same challenges faced by the multiple-choice format more generally. An interesting challenge for the explanation dialog model involves designing more open-ended interfaces that still support simple and reliable analysis of responses by the underlying software. Essentially, creating open-ended interfaces and formats is relatively easy—the challenge involves designing open-ended interfaces that elicit input that is easily and reliably analyzed to determine feedback, evaluation, and subsequent challenge difficulties. Section 5 of Table 10.2 outlines some of our initial ideas in terms of adapting free-body diagrams and concept maps toward this purpose, but ongoing work will be required to explore this limitation/challenge.

A related challenge involves selecting subject matter domain with multiple conceptually appropriate core ideas that players can leverage as focal "warrants" or explanations for their proposed solutions. In the example game discussed in this chapter, which focuses on mechanics, Newton's laws provide a relatively ideal set of concepts for this purpose. Students can explore and distinguish between the implications of the three laws as they invoke the laws at different times to explain different phenomena. For other domains, however, the underlying warrants for solutions might prove less conceptually rich and/or less well-aligned with the targeted concepts of the formal curriculum. The generalizability of the explanation dialog model therefore requires further exploration across other domains.

In terms of trade-offs and alternatives, at the most proximal level, we might replace IRT or HMM with other approaches within the explanation dialog model. HMM could be replaced, for example, with Sequential Pattern Analysis (Agrawal & Srikant, 1995; Zhou, Xu, Nesbit, & Winne, 2010). Sequential Pattern Analysis would be much less complicated to set up and would be computationally less expensive in terms of processing demands, but Sequential Pattern Analysis only analyzes sequences of events and not actual timings of events. Sequential Pattern Analysis therefore cannot distinguish between two actions taken sequentially vs. two actions separated by time. Therefore, Sequential Pattern Analysis might or might not make sense for specific applications of the explanation dialog model depending on the importance of timing considerations or other factors that one approach handled more or less effectively than the other.

Similarly, the affordances and limitations of the explanation dialog model itself should be compared with other game-based options depending on the characteristics of the underlying game to be assessed. Shute's stealth-based assessment offers another excellent model (Shute & Ke, 2012; Shute & Kim, in press). Stealth-based assessment is more broadly applicable to a larger number of game contexts and is more flexible in terms of what it might track. Training Bayesian nets for each challenge is substantially more complex, however, and connecting player actions to specific reactions or feedback by the game would be less precise. Similarly, SAVE Science's focus on making the core game about engaging in inquiry provides a more detailed assessment of students' ability to engage in inquiry (Nelson, Ketelhut, & Schifter, 2010; Schifter, Ketelhut, & Nelson, 2012), but the save science model is less flexible in terms of the breadth of game contexts to which it applies. As with any assessment choice, therefore, selection of the explanation dialog model, and its constituent components, should be a function of comparing affordances and limitations in light of other options and in light of the characteristics of the game context to which it will be applied.

10.10 Final Thoughts: Can the Explanation Dialog Model Be Fun?

The model for supporting learning games that we propose in this chapter depends on creating an "explanation game" that is central and fun in its own right on equal footing with the "core game." The model we have proposed leverages principles of game design that we believe will be effective at fostering engagement, while feeling familiar and comfortable to all learners. We will do so by generating a seamless flow of game play experience that interweaves levels of the core game with explanatory dialogs for formative assessment.

Engaging in dialog with simulated characters in the game world is a mechanism that has persisted since the earliest days of computer gaming. Since 1967, when Joseph Weizenbaum created *ELIZA*, a computer program designed to emulate interaction between the user and an artificial therapist, designers of interactive entertainment have attempted to incorporate meaningful interactions with virtual characters in order to aid immersion. While other conventions and genres have fallen into disuse, NPCs with branching dialog options remain key features of many game genres (e.g., role-playing and adventure games).

The interaction with the NPC is also important for a deeper reason: this narrative frame allows students to participate in help-giving, which not only provides a layer of meaning to play (cf. Gee, 2004; McGonigal, 2011), but has been demonstrated to be a key behavioral element for learning in small groups (Webb, 1989; Webb, Farivar, & Mastergeorge, 2002). To make the explanation game as similar as possible to the action of giving help to peers in a small group, we believe it is crucial that the dialog with NPCs is flexible, adaptive, and responsive to the learner; it is precisely this adaptive functionality that provides a workable core for an assessment strategy in our proposed model.

In our vision, having an assessment component in a game does not necessarily detract from engagement. We find many possible design choices can be made so that the explanation game and the core game interact in ways that students will find interesting and compelling. For example, by succeeding in the explanation game, students would unlock bonus levels, special one-time boosts for their characters ("power-ups"), customization options for their in-game avatars, etc. Since these awards are represented mainly in the core game, we view these awards as powerful forms of feedback to encourage success in the explanation game. We also envision rewards for outstanding play in the core game that provide smaller, but still significant, boosts that are applicable in the explanation game, such as extra time per challenge stage.

We believe that games for learning will engage students most powerfully if we present engaging, thought-provoking games and avoid the inconsistency of experience caused by interruptions for the purpose of assessment. The goal of our model is that the explanation game and the core game are perceived as two interwoven activities and that by integrating the rewards of one game into the play of the other game, we believe that the experience will be seamless; students will not perceive an interrupt in play and may not even be aware of which part of the game contains the assessment.

References

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. In *Proceedings of the eleventh IEEE international conference on data engineering (ICDE)* (pp. 3–14). Taipei, Taiwan.

American Association for the Advancement of Science. (1993). *Benchmarks for scientific literacy*. New York: Oxford University Press.

Amory, A. (2006). Game object model version II: A theoretical framework for educational game development. *Educational Technology Research and Development, 55*(1), 51–77.

Anderson, J., & Barnett, G. M. (2011). Using video games to support pre-service elementary teachers learning of basic physics principles. *Journal of Science Education and Technology, 20*(4), 347–362.

Annetta, L. A., Minogue, J., Holmes, S. Y., & Cheng, M.-T. (2009). Investigating the impact of video games on high school students' engagement and learning about genetics. *Computers in Education, 53*(1), 74–85.

Azevedo, R. (2009). Theoretical, conceptual, methodological, and instructional issues in research on metacognition and self-regulated learning: A discussion. *Metacognition and Learning, 4*(1), 87–95.

Barab, S. A., Arici, A., & Jackson, C. (2005). Eat your vegetables and do your homework: A design based investigation of enjoyment and meaning in learning. *Educational Technology, 45*(1), 15–20.

Barab, S. A., Sadler, T., Heiselt, C., Hickey, D., & Zuiker, S. (2007). Relating narrative, inquiry, and inscriptions: A framework for socio-scientific inquiry. *Journal of Science Education and Technology, 16*(1), 59–82.

Barab, S. A., Scott, B., Siyahhan, S., Goldstone, R., Ingram-Goble, A., Zuiker, S., et al. (2009). Transformational play as a curricular scaffold: Using videogames to support science education. *Journal of Science Education and Technology, 18*, 305–320.

Barab, S. A., Zuiker, S., Warren, S., Hickey, D., Ingram-Goble, A., Kwon, E.-J., et al. (2007). Situationally embodied curriculum: Relating formalisms and contexts. *Science Education, 91*(5), 750–782.

Bargh, J. A., & Schul, Y. (1980). On the cognitive benefits of teaching. *Journal of Educational Psychology, 72*(5), 593–604.

Bielaczyc, K., Pirolli, P., & Brown, A. L. (1995). Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem solving. *Cognition and Instruction, 13*(2), 221–252.

Biswas, G., Jeong, H., Kinnebrew, J., Sulcer, B., & Roscoe, R. (2010). Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Research and Practice in Technology Enhanced Learning, 5*(2), 123–152.

Biswas, G., Leelawong, K., Schwartz, D., Vye, N., & The Teachable Agents Group at Vanderbilt. (2005). Learning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence, 19*, 363–392.

Biswas, G., Schwartz, D., Bransford, J., & The Teachable Agents Group at Vanderbilt (TAG-V). (2001). Technology support for complex problem solving: From SAD environments to AI. In K. D. Forbus & P. J. Feltovich (Eds.), *Smart machines in education* (pp. 71–98). Menlo Park, CA: AAAI Press.

Borges, A. T., Tecnico, C., & Gilbert, J. K. (1998). Models of magnetism. *International Journal of Science Education, 20*(3), 361.

Brand, M., Oliver, N., & Pentland, A. (1997). Coupled hidden Markov models for complex action recognition. In *IEEE conference on computer vision & pattern recognition (CVPR)* (pp. 994–999). San Juan, Puerto Rico, June 17–19, 1997.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education, 24*, 61–100.

Caillois, R. (1961). *Man, play, and games* (1st U.S. ed.). New York: Free Press of Glencoe.

Champagne, A. B., Klopfer, L. E., & Gunstone, R. F. (1982). Cognitive research and the design of science instruction. *Educational Psychologist, 17*(1), 31.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*(2), 145–182.

Chi, M. T. H., & VanLehn, K. A. (1991). The content of physics self-explanations. *The Journal of the Learning Sciences, 1*(1), 69–106.

Clark, D. B. (2004). Hands-on investigation in internet environments: Teaching thermal equilibrium. In M. C. Linn, E. A. Davis, & P. Bell (Eds.), *Internet Environments for Science Education* (pp. 175–200). Mahwah, NJ: Lawrence Erlbaum Associates.

Clark, D. B. (2006). Longitudinal conceptual change in students' understanding of thermal equilibrium: An examination of the process of conceptual restructuring. *Cognition and Instruction, 24*(4), 467–563.

Clark, D. B., & Linn, M. C. (2003). Scaffolding knowledge integration through curricular depth. *The Journal of the Learning Sciences, 12*(4), 451–494.

Clark, D. B., & Sampson, V. D. (2007). Personally-seeded discussions to scaffold online argumentation. *International Journal of Science Education, 29*(3), 253–277.

Clark, D. B., D'Angelo, C., & Schleigh, S. (2011). Multinational comparison of students' knowledge structure coherence. *The Journal of the Learning Sciences, 20*(20), 207–261.

Clark, D. B., Nelson, B., Chang, H., D'Angelo, C. M., Slack, K., & Martinez-Garza, M. (2011). Exploring Newtonian mechanics in a conceptually-integrated digital game: Comparison of learning and affective outcomes for students in Taiwan and the United States. *Computers & Education, 57*(3), 2178–2195.

Clark, D. B., Nelson, B., D'Angelo, C. M., & Menekse, M. (2009). Integrating critique to support learning about physics in video games. In *Poster presented as part of a structured session at the National Association of Research in Science Teaching (NARST) 2009 meeting*, Garden Grove, CA.

Clark, D. B., Nelson, B., Martinez-Garza, M., & D'Angelo, C. M. (submitted). Digital games and science learning: Research across the NRC strands of science proficiency.

Clark, D. B., Nelson, B., Sengupta, P., & D'Angelo, C. M. (2009). Rethinking science learning through digital games and simulations: Genres, examples, and evidence. In *Invited topic paper in the proceedings of the national academies board on science education workshop on learning science: Computer games, simulations, and education*, Washington, DC.

Clarke, J., & Dede, C. (2005). Making learning meaningful: An exploratory study of using multiuser environments (MUVEs) in middle school science. In *Paper presented at the American Educational Research Association conference*, Montreal, Canada.

Coller, B., & Scott, M. (2009). Effectiveness of using a video game to teach a course in mechanical engineering. *Computers in Education, 53*(3), 900–912.

Csikszentmihalyi, M. (1991). Flow: The psychology of optimal experience. New York: Harper & Row, Publishers.

Dede, C., & Ketelhut, D. J. (2003). Designing for motivation and usability in a museum-based multi-user virtual environment. In *Paper presented at the American Educational Research Association conference*, Chicago, IL.

Dieterle, E. (2009). Neomillennial learning styles and River City. *Children, Youth and Environments, 19*(1), 245–278.

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction, 10*(2 & 3), 105–225.

diSessa, A. A. (1996). What do "just plain folk" know about physics? In D. R. Olson & N. Torrance (Eds.), *The handbook of education and human development: New models of learning, teaching, and schooling* (pp. 709–730). Oxford, UK: Blackwell Publishers.

Eignor, D. R., Way, W. D., Stocking, M. L., & Steffen, M. (1993). *Case studies in computer adaptive test design through simulation (research report # 93-56)*. Princeton, NJ: Educational Testing Service.

Federation of American Scientists. (2006). *Report: Summit on educational games: Harnessing the power of video games for learning*. Washington, DC: Federation of American Scientists.

Galas, C. (2006). Why Whyville? *Learning and Leading with Technology, 34*(6), 30–33.

Games-to-Teach Team. (2003). Design principles of next-generation digital gaming for education. *Educational Technology, 43*(5), 17–33.

Gee, J. P. (2003/2007). *What video games have to teach us about learning and literacy*. New York: Palgrave Macmillan.

Gee, J. P. (2004). *Situated language and learning: A critique of traditional schooling*. London: Routledge.

Gee, J. P. (2007). *Good video games and good learning: Collected essays on video games, learning and literacy (new literacies and digital epistemologies)*. New York: Peter Lang Publishing Inc.

Grant, P., Johnson, L., & Sanders, Y. (1990). *Better links: Teaching strategies in the science classroom*. Australia: STAV Publication.

Gunter, G., Kenny, R., & Vick, E. (2008). Taking educational games seriously: Using the RETAIN model to design endogenous fantasy into standalone educational games. *Educational Technology Research and Development, 56*(5), 511–537. doi:10.1007/s11423-007-9073-2.

Hammer, D., Elby, A., Scherr, R. E., & Redish, E. F. (2005). Resources, framing, and transfer. In J. P. Mestre (Ed.), *Transfer of learning from a multidisciplinary perspective* (pp. 89–119). Greenwich, CT: Information Age Publishing.

Hickey, D., Ingram-Goble, A., & Jameson, E. (2009). Designing assessments and assessing designs in virtual educational environments. *Journal of Science Education and Technology, 18*(2), 187–208.

Hines, P. J., Jasny, B. R., & Merris, J. (2009). Adding a T to the three R's. *Science, 323*, 53.

Holbert, N. (2009). Learning Newton while crashing cars. In *Poster presented at games, learning and society*, Madison, WI, June 10–12, 2009.

Honey, M. A., & Hilton, M. (Eds.). (2010). *Learning science through computer games and simulations. National Research Council*. Washington, DC: National Academy Press.

Huizinga, J. (1980). *Homo Ludens: A study of the play element in culture*. London: Routledge and Kegan.

Hunt, E., & Minstrell, J. (1994). A cognitive approach to the teaching of physics. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 51–74). Cambridge, MA: MIT Press.

Jeong, H., & Biswas, G. (2008). Mining student behavior models in learning-by-teaching environments. In *Proceedings of the first international conference on educational data mining* (pp. 127–136). Montreal, Canada.

Juang, B. H., & Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics, 33*(3), 251–272.

Kafai, Y. B., Quintero, M., & Feldon, D. (2010). Investigating the 'why' in Whypox: Casual and systematic explorations of a virtual epidemic. *Games and Culture, 5*(1), 116–135.

Kearney, M. (2004). Classroom use of multimedia-supported predict–observe–explain tasks in a social constructivist learning environment. *Research in Science Education, 34*(4), 427–453.

Kearney, M., & Treagust, D. (2000). An investigation of the classroom use of prediction-observation-explanation computer tasks designed to elicit and promote discussion of students' conceptions of force and motion. In *Presented at the national association for research in science teaching*, New Orleans, USA.

Ketelhut, D. J., Dede, C., Clarke, J., & Nelson, B. (2006). A multi-user virtual environment for building higher order inquiry skills in science. In *American Educational Research Association conference,* San Francisco, CA.

Kinnebrew, J. S., Loretz, K. M., & Biswas, G. (in press). A contextualized, differential sequence mining method to derive students' learning behavior patterns. *Journal of Educational Data Mining*, 2012.

Kinnebrew, J. S., Loretz, K. M., & Biswas, G. (2011). Modeling and measuring self-regulated learning in teachable agent environments. *Journal of e-Learning and Knowledge Society, 7*(2), 19–35.

Klopfer, E., Scheintaub, H., Huang, W., Wendal, D., & Roque, R. (2009). The simulation cycle: Combining games, simulations, engineering and science using StarLogo TNG. *E-learning, 6*(1), 71–96.

Krogh, A., Brown, M., Mian, S., Sjolander, K., & Haussler, D. (1994). Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology, 235*(5), 1501–1531.

Leelawong, K., & Biswas, G. (2008). Designing learning by teaching agents: The Betty's brain system. *International Journal of Artificial Intelligence in Education, 18*(3), 181–208.

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice, 26*(2), 3–16.

Li, C., & Biswas, G. (2002). Applying the hidden Markov model methodology for unsupervised learning of temporal data. *International Journal of Knowledge Based Intelligent Engineering Systems, 6*(3), 152–160.

Liew, C. W., & Treagust, D. F. (1995). A predict-observe-explain teaching sequence for learning about students' understanding of heat and expansion liquids. *Australian Science Teachers Journal, 41*(1), 68–71.

Liew, C. W., & Treagust, D. F. (1998). The effectiveness of predict-observe-explain tasks in diagnosing students' understanding of science and in identifying their levels of achievement. In *Presented at the American Educational Research Association*, San Diego, CA.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum Associates.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement, 20*, 389–404.

Masson, M. E. J., Bub, D. N., & Lalonde, C. E. (2011). Video-game training and naive reasoning about object motion. *Applied Cognitive Psychology, 25*(1), 166–173.

Mayer, R. E., & Johnson, C. I. (2010). Adding instructional features that promote learning in a game-like environment. *Journal of Educational Computing Research, 42*(3), 241–265.

Mazur, E. (1996). *Peer instruction: A user's manual (Pap/Dskt)*. San Francisco, CA: Benjamin Cummings.

McGonigal, J. (2011). *Reality is broken: Why games make us better and how they can change the world*. New York: Penguin Press.

McQuiggan, S., Rowe, J., & Lester, J. (2008). The effects of empathetic virtual characters on presence in narrative-centered learning environments. In *Proceedings of the 2008 SIGCHI conference on human factors in computing systems* (pp. 1511–1520), Florence, Italy.

Minstrell, J. (1982). Explaining the "at rest" condition of an object. *The Physics Teacher, 20*, 10–14.

Minstrell, J. (1989). Teaching science for understanding. In L. Resnick & L. Klopfer (Eds.), *Toward the thinking curriculum* (pp. 129–149). Alexandria, VA: Association for Supervision and Curriculum Development.

Minstrell, J., & Kraus, P. (2005). Guided inquiry in the science classroom. In M. S. Donovan & J. D. Bransford (Eds.), *How students learn: History, mathematics, and science in the classroom*. Washington, DC: National Academies Press.

Moreno, R., & Mayer, R. E. (2000). Engaging students in active learning: The case for personalized multimedia messages. *Journal of Educational Psychology, 92*, 724–733.

Moreno, R., & Mayer, R. E. (2004). Personalized messages that promote science learning in virtual environments. *Journal of Educational Psychology, 96*, 165–173.

National Research Council. (1996). *The national science education standards*. Washington, DC: The National Academy Press.

National Research Council. (2010). In M. Hilton (Ed.), *Exploring the intersection of science education and 21st century skills: A workshop summary*. Washington, DC: National Academy Press.

National Research Council. (2012). *Conceptual framework for new science education standards*. Washington, DC: National Academy of Sciences Board on Science Education.

Nelson, B. (2007). Exploring the use of individualized, reflective guidance in an educational multiuser virtual environment. *Journal of Science Education and Technology, 16*(1), 83–97.

Nelson, B., Ketelhut, D., Clarke, J., Bowman, C., & Dede, C. (2005). Design-based research strategies for developing a scientific inquiry curriculum in a multi-user virtual environment. *Educational Technology, 45*(1), 21–34.

Nelson, B., Ketelhut, D. J., & Schifter, C. (2010). Exploring cognitive load in immersive educational games: The SAVE science project. *International Journal for Gaming and Computer Mediated Simulations, 2*(1), 31–39.

Neulight, N., Kafai, Y. B., Kao, L., Foley, B., & Galas, C. (2007). Children's participation in a virtual epidemic in the science classroom: Making connections to natural infectious diseases. *Journal of Science Education and Technology, 16*(1), 47–58.

Palmer, D. (1995). The POE in the primary school: An evaluation. *Research in Science Education, 25*(3), 323–332.

Parnafes, O., & diSessa, A. A. (2004). Relations between types of reasoning and computational representations. *International Journal of Computers for Mathematical Learning, 9*, 251–280.

Pelletier, C. (2008). Gaming in context: How young people construct their gendered identities in playing and making games. In Y. B. Kafai, C. Heeter, J. Denner, & J. Y. Sun (Eds.), *Beyond Barbie and Mortal Kombat: New perspectives on gender and gaming* (pp. 145–158). Cambridge, MA: The MIT Press.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451–502). San Diego: Academic.

Quellmalz, E. S., & Pellegrino, J. W. (2009). Technology and testing. *Science, 323*(5910), 75–79.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.

Roy, M., & Chi, M. T. H. (2005). The self-explanation principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 271–286). New York: Cambridge University Press.

Salen, K., & Zimmerman, E. (2003). *Rules of play: Game design fundamentals (illustrated edition)*. Cambridge, MA: The MIT Press.

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

Schifter, C. C., Ketelhut, D. J., & Nelson, B. C. (2012). Presence and middle school students' participation in a virtual game environment to assess science inquiry. *Educational Technology & Society, 15*(1), 53–63.

Schraw, G., Crippen, K., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education, 36*(1), 111–139.

Schwartz, D. L., Blair, K. P., Biswas, G., & Leelawong, K. (2007). Animations of thought: Interactivity in the teachable agent paradigm. In R. Lowe & W. Schnotz (Eds.), *Learning with animation: Research and implications for design* (pp. 114–140). Cambridge, UK: Cambridge University Press.

Schwartz, D. L., Chase, C., Chin, C., Oppezzo, M., Kwong, H., Okita, S., et al. (2009). Interactive metacognition: Monitoring and regulating a teachable agent. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education*. New York: Routledge Press.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction, 22*(2), 129–184.

Scott, P. H., Asoko, H. M., & Driver, R. H. (1991). Teaching for conceptual change: A review of strategies. In R. Duit, F. Goldberg, & H. Niederer (Eds.), *Research in physics learning: Theoretical issues and empirical studies* (pp. 310–329). Kiel, Germany: Schmidt & Klannig.

Searle, P., & Gunstone, R. (1990). Conceptual change and physics instruction: A longitudinal study. In *Presented at the American Educational Research Association*, Boston, MA.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*(2), 331–354.

Sengupta, P. (2011). Learning electromagnetism with ElectroHub—A digital game based on participatory simulation. Digital games and science learning. In D. Clark (Org.), *Invited paper session at the Annual Conference of National Association of Research on Science Teaching* (NARST 2011) Orlando, FL.

Sengupta, P., & Wilensky, U. (2009). Agent-based models and learning electricity. In *Paper presented at the annual meeting of the American Educational Research Association (AERA 2009)*, New York, NY.

Sengupta, P., & Wilensky, U. (2011). Lowering the learning threshold: Multi-agent-based models and learning electricity. In M. S. Khine & I. M. Saleh (Eds.), *Dynamic modeling: Cognitive tool for scientific inquiry* (pp. 141–171). New York: Springer.

Shepardson, D. P., Moje, E. B., & Kennard-McClelland, A. M. (1994). The impact of a science demonstration on children's understandings of air pressure. *Journal of Research in Science Teaching, 31*(3), 243–258.

Shute, V. J., & Ke, F. (2012). Games, learning, and assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives*. New York, NY: Springer.

Shute, V. J., & Kim, Y. J. (in press). Formative and stealth assessment. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (4th ed.). New York, NY: Lawrence Erlbaum Associates, Taylor & Francis Group.

Shute, V. J., Rieber, L., & Van Eck, R. (2011). Games… and… learning. In R. Reiser & J. Dempsey (Eds.), *Trends and issues in instructional design and technology* (3rd ed., pp. 321–332). Upper Saddle River, NJ: Pearson Education, Inc.

Squire, K. (2005). Changing the game: What happens when video games enter the classroom. *Innovate, 1*(6), 25–49.

Squire, K., Barnett, M., Grant, J. M., & Higginbotham, T. (2004). Electromagnetism supercharged!: Learning physics with digital simulation games. In Y. B. Kafai, W. A. Sandoval, N. Enyedy, A. S. Nixon, & F. Herrera (Eds.), *Proceedings of the 6th international conference on learning sciences* (pp. 513–520). Los Angeles: UCLA Press.

Squire, K., & Jan, M. (2007). Mad City Mystery: Developing scientific argumentation skills with a place-based augmented reality game on handheld computers. *Journal of Science Education and Technology, 16*(1), 5–29.

Squire, K., & Klopfer, E. (2007). Augmented reality simulations on handheld computers. *The Journal of the Learning Sciences, 16*(3), 371–413.

Steinkuehler, C., & Duncan, S. (2008). Scientific habits of mind in virtual worlds. *Journal of Science Education and Technology, 17*(6), 530–543.

Tao, P., & Gunstone, R. F. (1999). The process of conceptual change in force and motion during computer-supported physics instruction. *Journal of Research in Science Teaching, 36*(7), 859–882.

Van der Linden, W., & Glas, C. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika, 75*(1), 120–139.

Webb, N. M. (1989). Peer interaction and learning in small groups. *International Journal of Educational Research, 13*(1), 21–39.

Webb, N. M., Farivar, S. H., & Mastergeorge, A. M. (2002). Productive helping in cooperative groups. *Theory into Practice, 41*(1), 13–20.

White, B. C., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*(1), 3–117.

White, B. C., & Frederiksen, J. R. (2000). Technological tools and instructional approaches for making scientific inquiry accessible to all. In M. J. Jacobson & R. B. Kozma (Eds.), *Innovations in science and mathematics education* (pp. 321–359). Mahwah, NJ: Lawrence Erlbaum Associates.

White, R. T., & Gunstone, R. F. (1992). *Probing understanding*. New York: Routledge.

Winne, P., & Hadwin, A. (2008). The weave of motivation and self-regulated learning. In D. Schunk & B. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory, research, and applications* (pp. 297–314). New York: Taylor & Francis.

Wright, W. (2006). Dream machines. *Wired, 14*(4), 110–112.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Washington, DC: American Council on Education/Praeger.

Zhou, M., Xu, Y., Nesbit, J. C., & Winne, P. H. (2010). Sequential pattern analysis of learning logs: Methodology and applications. In C. Romero (Ed.), *Handbook of educational data mining* (p. 107). Boca Raton, FL: CRC Press.

Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: An overview and analysis. In B. Zimmerman & D. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (pp. 1–37). Mawah, NJ: Erlbaum.