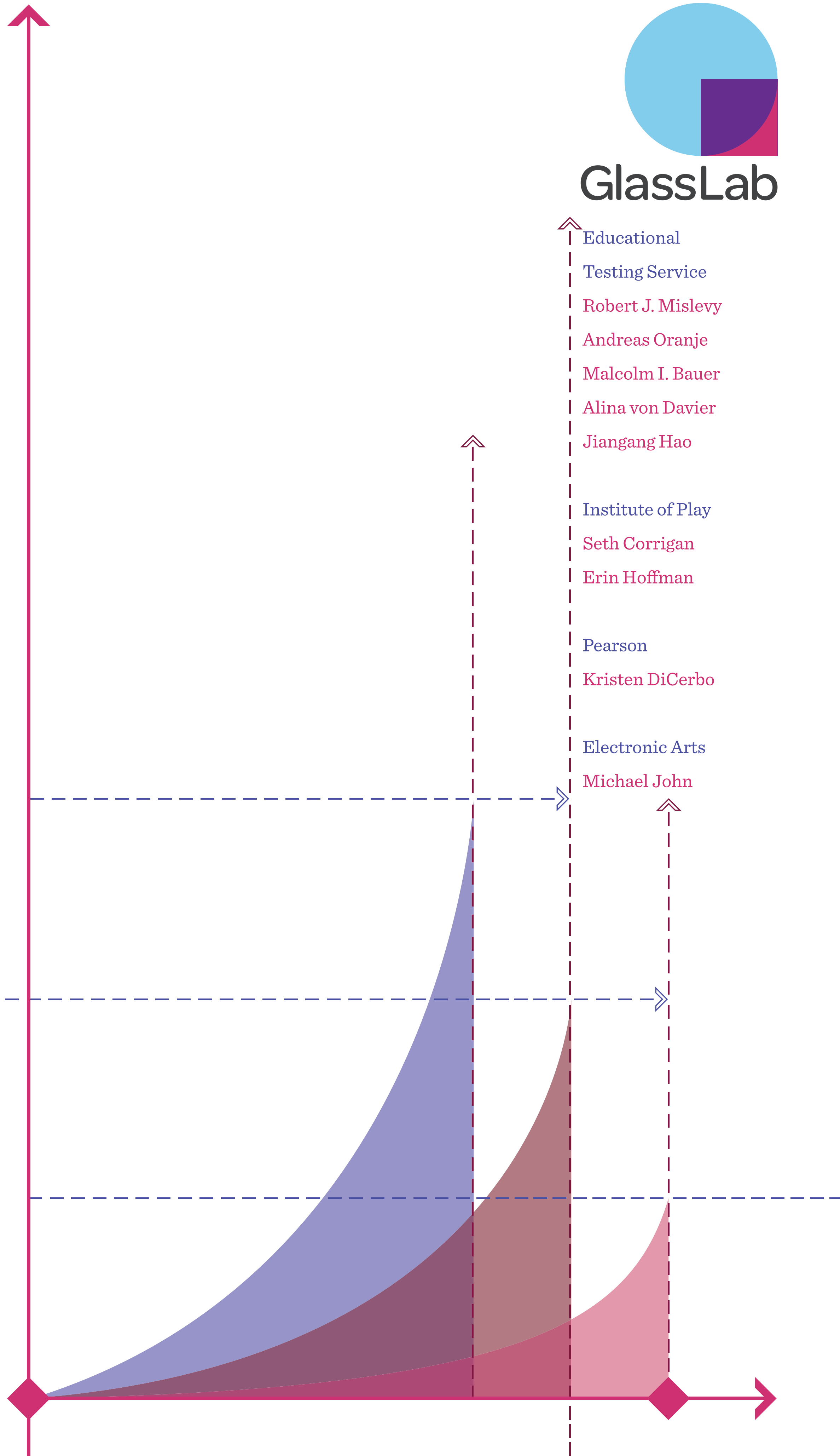
Psychometric Considerations In Game-Based Assessment

GlassLab



Educational Testing Service

Robert J. Mislevy

Andreas Oranje

Malcolm I. Bauer

Alina von Davier

Jiangang Hao

Institute of Play

Seth Corrigan

Erin Hoffman

Pearson

Kristen DiCerbo

Electronic Arts

Michael John

Foreword

In 1954, two analysts from the Rand Corporation presented a paper on simulation games as part of a symposium on the use and value of games to war methods. "Gaming as a Technique of Analysis," (Mood and Specht, 1954) argued for gaming as a strategy for discovering optimal choice within a system of complex possible outcomes. Focusing on the human qualities of judgment and intuition, the paper describes the way the process of decision-making can be modeled by a digital game, "...a black box into which we crank inputs and out of which are ground outputs." While we might not naturally think about games as machines—they hardly seem machine-like in their spontaneous and improvisational expression of play—games can be understood as state machines, or models of behavior composed of states, transitions, and actions. As game designer Warren Robinett points out, "A video game is a simulation, a model, a metaphor" (Robinett, 2005, p690). This definition of games as models is important, for it points to their status as artificial systems, systems that reflect the values and expertise of their designers. As designed models, games embed man in both their creation and in their play.

To analysts of the 1950s, gaming provided an observable and repeatable system where multiple scenarios could be quantitatively assessed and tested by players who, despite their fallibility as humans, carry with them the power of creative thought, intuition, and speculation. These players are bound by the rules of the game, act within these constraints, and tend to optimize their choices in pursuit of the best possible outcome. The rules never solely determine the play of a game; they are always set into motion by players with their own wants, skills, and expectations. Then, as now, this power is what sets gaming apart from pure machinic calculation. "To sit down and play through a game is to be convinced as by no argument, however persuasively presented" (Mood and Specht, 1954).

"Psychometric Considerations in Game-based Assessment" continues this tradition of inquiry into the use of games as models with a "human decision link," or games whose effectiveness is measured not only by the workings of computer code, but also by the actions of players. And, in the case of the simulation games discussed in this paper, the actions of student learners. Bringing the concepts and techniques of assessment and psychometrics to bear on the problem of game-based learning, the authors lay out an explicit framework linking the concerns and practices of game development to that of assessment design.

A primer, of sorts, for the emerging field of game-based assessment, the paper focuses specifically on the formative assessment value of simulation games, and in particular, the game SimCityEDU: Pollution Challenge!. Using the game as a case study, the paper explores the ways in which psychometric considerations specifically, and assessment design more generally, can be integrated into the game development process. And while the terminology can get quite technical at times, just keep in mind the authors' larger goal of pinpointing the unique overlap between the mechanics of games, assessment, and learning. They, like the analysts from an earlier era, believe in the power of games to not only engage, but also to reveal something about the capabilities of the learners engaged in their play.

This paper is the first of several papers to be published by Institute of Play on the work and research of GlassLab. The second paper, tentatively titled "Developing Game-based Assessments: An Agenda For Research And Design," is scheduled for release in Fall 2014, with a third coming in Summer 2015. You can follow the work of GlassLab at www.glasslabgames.org.

Katie Salen
Principal, Institute of Play
Game Designer

Table of Contents

Figures

Preface

This book is a companion to "Three things game designers need to know about assessment" (Mislevy, Behrens, DiCerbo, Frezzo, & West, 2012), which, by the way, are:

- The principles of assessment design are compatible with the principles of game design. It is because they both build on the same principles of learning.
- Assessment isn't really about numbers; it's about the structure of reasoning.[1]
- The key constraints of assessment design and game design need to be addressed from the very beginning of the design process.

We discuss here what assessment designers and psychometricians need to know about game-based assessment—more broadly, how to think about a given game-based assessment (GBA) from the perspective of a psychometrician, but integrated with the key ideas and goals of other domains that are fundamentally important to its success. The first half of the piece doesn't look like something out of Psychometrika or the Journal of Educational Measurement. We need to return to fundamentals of learning and design, of assessment arguments and evidentiary reasoning, to understand when and how the underlying concepts of psychometrics can be useful in GBA. We can then see how to integrate the concepts into GBA design from the beginning, and apply, adapt, or invent machinery to put them to work.

Chapters 1-5 are background that is meant to be broadly accessible, to game designers, subject-area experts, and learning scientists as well as measurement specialists. Chapter 6, on assessment design, is pivotal: It provides a conceptual framework for assessment design, which at once connects the game and learning aspects of a GBA with the assessment aspects, and gives meaning to psychometric modeling that may follow. Chapters 7-12 are more aimed at the measurement specialist, especially Chapter 10. We have tried to make these chapters readable, and we hope useful, to motivated readers from the allied fields. Chapter 13 discusses implications for GBA design, drawing on practices from the game industry, instructional science, and assessment design, and our own experience in trying to integrate them.

[1] Pearl (1988) quoted the statistician Glenn Shafer as having said, "Probability isn't really about numbers; it's about the structure of reasoning."

Introduction

Advances in technology and learning science open the door to a radically new vision of learning and assessment, characterized by the interaction and adaptation that digital environments afford. Game and simulation environments in particular provide students opportunities to develop and demonstrate proficiencies in complex interactive situations (Klopfer, Osterweil, & Salen, 2009). Naturally, designing effective and valid systems raises many challenges: Getting the content right, targeting the right levels of skills, keeping students engaged, and providing useful information to students, teachers, and the system itself as it interacts with students. The last of these, providing useful information, is a matter of reasoning from the specific things that students do, to what they know and do more broadly, and what the system, the teacher, or the students themselves might do next to develop their capabilities further. This book explores how the ideas and methods of psychometrics can contribute to this challenge in game-based assessment.

The goal of assessment is to gather and make sense of information about what students know and can do, for some purpose; evaluating their progress, for example, or shaping their next learning experience. The information will be better to the degree that students are engaged with the experience, and putting forth effort (Schmit & Ryan, 1992; Sundre & Wise, 2003). Games have the capability to engage students and to create conditions that foster learning (Gee, 2007). The hypothesis behind game-based assessment (GBA for short) is that GBA may offer a sweet spot in the assessment design space for some purposes and some circumstances. We will sketch a variety of ways GBA might be used for different purposes, by different users (that is, "use cases"). Our focus, though, will be on uses that seem particularly suited to the strengths of games, namely formative assessment to guide learning within a simulation environment. This means we will need to draw on ideas from instructional design as well as game design and assessment design.

Why Psychometrics?

The word psychometrics means "mental measurement." It originated more than a century ago with an aim of measuring traits—but the models don't know this. In educational assessment, psychometricians and statisticians have developed a toolkit to support reasoning from noisy data in real-world problems, to help monitor and guide learning. There are concepts and techniques for gathering information about what people know and can do, and methods for characterizing the amount and quality of evidence for given purpose. Our concern lies in this reasoning-about-evidence aspect of psychometrics. Quantitative methods for reasoning about evidence do not require any presumption of quantitative traits "inside people's heads."

Much of the machinery and activity in games looks quite different from familiar assessments, but fundamental issues of evidence in design and analysis arise in GBA as they do in any assessment. Particular challenges for psychometrics in GBA can include multiple, interacting, aspects of knowledge and skill; construct-irrelevant variation from game features; dependencies among actions across time points; different situations arising for different players as they interact with a game; and the fact that "tasks" and salient features of performance need not be predefined or the same across students.

As we begin explore these issues in GBA, questions arise naturally: Why would we want to do this? Don't good designers already design games that help students learn? What is the value-added, if any, from all that psychometric machinery? It might be useful in high stakes tests, but is it relevant to formative tests, where stakes are low, the game is just meant to support learning, and there are lots of opportunities to adjust along the way? In short, why psychometrics?

Good questions all. These are our reasons for investigating psychometrics in game-based assessment.

Reason #1: Psychometrics works together with assessment design to provide an explicit design framework for evoking evidence.

Getting students to think and act in the ways that are central to the targeted learning is at the heart of GBA. That's why instructional design is useful in GBA. Good insights of this kind can motivate the targeted thinking in students. But evoking the thinking is not the same as evoking evidence about the thinking. Assessment design and psychometrics together are about reasoning from students' actions to capabilities. What are the features of situations that evoke students' thinking? But then what can we learn about their thinking from what they actually do? Why is it that these particular kinds of situations give us evidence about what we care about? How do we identify and synthesize many and varied clues? Can a student win a GBA game without learning anything or without being good at the thing we want them to learn? Can she learn but not win? Some of the tools of educational measurement we will use work well because they developed in environments where people sue you if you don't have good answers to questions like these. We want to learn how to use them in conjunction with game-design tools ... which are good for an equally compelling reason: If a game isn't engaging, you quit playing, and you tell your friends not to play it.

Reason #2: Psychometrics provides an explicit framework for characterizing evidence.

Psychometrics provides both qualitative and quantitative methods for talking about, then characterizing, principles like reliability, validity, and comparability when we reason from limited evidence. It is true that intuitive, informal reasoning from evidence can be quite satisfactory in many assessment situations, especially when the stakes are low. And, truth be known, getting the right kinds of evidence about the right learning is more important than managing evidence efficiently. But intuitive methods for dealing with evidence don't tell us about the qualities of the evidence. Psychometric machinery provides a meta framework for assessment—a framework for characterizing evidence about evidence. An insightful designer's GBA might indeed turn out to provide excellent evidence about learning. But a psychometric framework is needed to provide evidence about its evidence about learning.

Reason #3: We need to characterize evidence for moderate or higher stakes purposes.

When a GBA is helping students learn about the right kinds of things and assessment information is being used locally, the specifics of the quality of evidence are not so critical. It is easier for the student or the teacher or the GBA itself to see when something is wrong, and change. When the information will be used by someone outside that local setting and be taken seriously in its own right, we need to be more concerned about its qualities as evidence. This is so when we begin to talk about grades, badges, or credit. All the more if the information is part of a decision for graduation, licensure, or program evaluation. The psychometric framework helps us examine things like the value of information, the quality of decisions, and tradeoffs of assessment time versus value of evidence (not simply a matter of how much, but about which kinds, in what balance, under what constraints).

Reason #4: We need more effective tools when evidentiary-reasoning problems get complicated.

Simply adding up scores works pretty well for combining multiple bits of similar kinds of evidence, for the purpose of determining how well somebody is doing at that sort of thing (although, as per Reason #2, we don't know yet how well). It is not so easy to sort out the evidence when different aspects of knowledge, or skill, or experience are called on in different ways in different situations; or when they are manifest in different aspects of performance; or when different parts of a challenge depend on other parts. Sometimes in a GBA we may need to unravel these evidentiary complexities, say to give feedback about different aspects of performance, or decide what kind of challenge to pose next, or adjust the current challenge during play to become harder in one respect but less demanding in another. Psychometric models can help us do this (Chapter 10).

Reason #5: Psychometrics provides metrics to improve design with respect to evidence.

Game designers improve design by lots of testing--early, often, and continually. What parts arefun? Where do people get stuck? Which features confuse them, and which ones motivate them? Key tools are to improve play are A/B tests and metrics. An A/B test is an experiment. In its basic form, the A group of players and the B group play games that are similar except for some particular difference. Which way works better? Designers talk to the players, but they also look at metrics, such as questions about fun, measures of engagement, or statistics for completion and quitting. Knowing what you want to maximize allows you to compare designs to get it. Designers can do this with a GBA to tune its game play aspects. Metrics for evidence will similarly enable them to tune its evidentiary aspects: What features in the game situation improve or decrease evidence? Which mechanic provides better evidence? How much evidence about which aspects of proficiency do these two scenarios provide? Is there a better way to identify evidence from this stream of play?

Reason #6: Working out psychometrics for GBA advances psychometrics.

Most psychometric methods for educational assessment apply to what DiCerbo and Behrens (2012) called "the digital desert": relatively sparse, self-contained, bits of evidence gleaned in familiar kinds of assessment such as answers to multiple-choice questions and raters' scores for students' essays. Digital environments can make available, in real time, every click, keystroke, and other interaction being recorded by the technology. Seemingly limitless data could be mined to inform and predict student learning, seamlessly embedded within naturalistic assessment and learning activities (Shute, 2011). The field of psychometrics is challenged to extend insights it has developed over the past century for reasoning from simpler forms of evidence, to now support reasoning in "the digital ocean." Game-based assessment is a critical arena for learning how to do it—critical for both psychometrics and for users. Educators are developing GBAs already, and are already using them to shape students' learning and to make educational decisions. But how well do they work? Users benefit if there are tools to examine the quality of these uses. Psychometricians, if they are to contribute to new forms of assessment, must figure out how to extend their tools to questions of evidence and inference in the digital ocean.

Reason #7: A principled design and analysis framework contributes to efficiency and validity in GBA design.

We stipulate as above that insightful designers can produce effective GBAs without formal assessment design and psychometric machinery. But not everyone is a gifted designer. An integrated game/assessment design framework would provide considerable advantage to "the rest of us." Examples of benefits include a shared language for game designers and assessment designers to tackle together their joint design problem as it plays out in each particular project; integrated design processes to help them do this more efficiently; and re-usable components that integrate game mechanics and evidence-bearing opportunities (Chapter 13).

The Design Challenge

Many of the challenges to measurement modeling in GBA also show up in performance testing and simulation-based tasks. We can draw on those literatures for insights. But a particular challenge of GBAs is that designers need to satisfy constraints and serve purposes beyond measurement. In particular, games are meant to be engaging—even fun—in ways that assessments usually are not. Designers of recreational games have developed successful practices and "mechanics" to engage players (Salen & Zimmerman, 2004), but generally not with the aim of supporting inferences about students' capabilities in substantive learning domains. A GBA must serve purposes beyond those in a game meant mainly for entertainment.

From the players' perspective, they are in a situation with some features, there are ways to act, and there are goals to pursue. From the GBA designers' perspective, we want at once to leverage game design principles and mechanics to generate engagement, and assessment design principles and methods to produce useful evidence. The game and assessment communities have little in the way of common vocabulary, shared principles, or joint game/assessment mechanics to design in the combined GBA space. Our discussion of how to integrate the diverse aspects of the design problem will draw on ideas from evidence-centered assessment design (ECD; Mislevy, Steinberg, & Almond, 2003), since they were developed to address such challenges, and explore how they can be integrated with the world of game design.

We aim to contribute to a common conceptual foundation for GBA design, which helps team members from across substantive, game, and psychometric specialties to work together towards their common goal. The emphasis here is the assessment design and psychometric modeling aspects, and this discussion is necessarily technical at points. We explore how evidence-handling psychometric machinery that evolved for assessment can be adapted to GBA, and show how psychometric considerations interact with considerations from the equally-involved design domains for games and learning. We will use several illustrations from a GBA called SimCityEDU that we developed in the GlassLab project, as well as occasional examples from commercial games that some readers will find helpful, and from simulation-based assessments and intelligent tutoring systems that share some design challenges with GBAs and offer insights for GBA design.

A central message is that applying psychometric concepts to GBA is not simply a matter of applying psychometric methods after-the-fact to games that have been optimized for learning and engagement, then "figuring out how to score them." A better design process jointly addresses the concerns of game design, instructional design, and assessment as required, so that key considerations of each perspective are taken into account from the beginning (Mislevy, Behrens, DiCerbo, Frezzo, & West, 2012). This integrated approach encourages designers to recognize trade-offs that cut across design domains and devise solutions that balance concerns across them.

In order to talk about psychometric considerations, we find we must lay out a fair amount of terminology and representations for assessment design. It is at the level of assessment design rather than psychometrics per se that integration with game design and instructional design must occur. An important component of the discussion will be to lay out ways that games, learning, and assessment share goals and perspectives, and where they differ and can provoke design tradeoffs.

Roadmap

Chapter 2 summarizes the psychometric mindset we bring to the presentation. There is a lot in here that doesn't look like psychometrics proper. Not many equations. Hardly anything about estimation algorithms or model fit. A lot more about learning and purposes, designs and arguments. And everything that's here about models and parameters is grounded in assessment arguments. This chapter tells why we think this is the most important part of the undertaking.

Chapter 3 does more stage setting for game-based assessment. It introduces Jackson City, a challenge from a GBA that we will use to illustrate ideas throughout the presentation. It then discusses where evidence in GBA can come from, and different ways that games can be used in connection with assessment. These so-called use cases can differ quite a bit from one another, with implications for design and for psychometric properties like validity and generalizability. Not all GBA use cases are equal.

Chapter 4 summarizes the sociocognitive perspective on learning that our approach to psychometrics in GBA is based on. It is not the psychological perspective under which psychometrics evolved. But it is a perspective that is well suited to the learning and game-play aspects of GBA—and, we believe, a perspective to which the concepts and methods of psychometrics can be usefully applied.

Chapter 5 reviews some of the concepts in game design that interact with assessment design and the meaning of the models and variables in psychometric models. These are important for assessment designers and psychometricians who are working on GBA design teams to understand, since design tradeoffs can cut across design domains, and understanding the problems and the tools of other design disciplines involved in a GBA helps team members work together to achieve the common goals.

Chapter 6 walks through the evidence-centered assessment design framework (ECD; Almond, Steinberg, & Mislevy, 2002; Mislevy, Steinberg, & Almond, 2003). This is where we lay out the logic and structure of assessment arguments, then how particular elements of it are instantiated in the situations and activities of an assessment—in particular, when that assessment is a GBA.

Chapter 7 provides a more grounded explanation of the role of psychometrics in GBA, drawing on the concepts and language of Chapter 6.

Chapters 8-10 dig more deeply into the more formally psychometric topics of work products, evidence identification (colloquially, scoring), and measurement models.

Chapter 11 discusses issues of meaning and use of psychometric modeling that are particularly important in game-based assessment, which differ from uses with more familiar kinds of assessments. These include the situated meanings of measurement-model variables in GBAs, changing values of latent variables as players learn, multiple plays, and collaboration.

Chapter 12 addresses the topic of the psychometric properties reliability, generalizability, comparability, and validity as they arise in game-based assessment. We argue that they are as important as they are in any other kind of assessment, but appearing in ways and analyzed with methods that are not always the same as for familiar kinds of assessment.

Chapter 13 focuses on implications of a psychometric perspective for the design of game-based assessment. We mentioned above that few if any individuals come to a GBA project as experts in learning, game design, and assessment design, let alone an integration of these domains. This chapter describes some strategies and representations we have found useful to this end, in a process we call Evidence-Centered game Design, or ECgD.

How We Think About Psychometrics for Game-Based Assessment

Psychometrics is often thought of narrowly as machinery -- the models and procedures per se. This can be a sufficient way to think about psychometrics for familiar kinds of assessments for familiar kinds of purposes. However, these models and machinery are functionally about reasoning with information and uncertainty in the limited, noisy, and context-bound observations in assessment. This kind of thinking is implicit in its uses with familiar models and procedures. It is intertwined with assumptions about the nature of proficiency, the properties of evidence, and the kinds of inferences and decisions that scores will be used to support.

Yet the underlying principles are just as relevant for new forms of assessment, where familiar models and methods don't always apply. To quote Samuel Messick (1994), "such basic assessment issues as validity, reliability, comparability, and fairness need to be uniformly addressed for all assessments because they are not just measurement principles, they are social values that have meaning and force outside of measurement wherever evaluative judgments and decisions are made" (p. 13). These principles apply no less in game-based assessment, in ways that are appropriate to the ways a GBA is being used. These may be quite different than they are in high-stakes tests or classroom quizzes, and may require different machinery to tackle them. To do so, we need to bring out beliefs about the nature of proficiency and evidence in new situations, new forms of data, and new uses. We need to determine, when necessary from first principles, the underlying webs of assumption and reasoning that support them; and we need to recognize where a psychometric perspective can help us build them, explicate them, critique them, and support practical work through them.

To adapt psychometric thinking to new kinds of assessment such as GBA, then, we need to be able to integrate psychometric thinking with thinking about learning, psychology, game design, and social embedding. Sometimes the models and procedures will be familiar ones, applied in familiar ways. Other times they will be similar methods, but re-interpreted for different contexts with analogous information-management characteristics. Still other times, extensions or new models may be required, extending the underlying principles. But whatever psychometrics is needed will need to be tuned with other (sometimes competing) features of a GBA--an artifact that spans design domains with respect to both purposes and techniques for achieving them.

The models and the methods of psychometrics will be used in GBA to synthesize evidence about aspects of students' activities and capabilities. As well as providing a basis for feedback and reporting, psychometric models provide ways to characterize the amount and quality of the evidence. The use the framework of probability-based reasoning to do so, with models for the relationships between aspects of students' (inherently unobservable) capabilities and the (observable) things they say and do in various situations. Much of the machinery of psychometrics evolved in the context of trait and behaviorist psychology, for performances in well-defined tasks. However, the same principles and models, appropriately conceived and extended, can be applied much more broadly -- for example, in simulation-based tasks (Mislevy, 2013), portfolio assessment (Wolfe & Gitomer, 2001), and data mining (Mislevy, Behrens, DiCerbo, & Levy, 2012).

These ideas are important in GBA because psychometrics provides a principled way to study the assessment aspect of a GBA: It provides metrics, for example, to know how accurate it is, and when there are design options, to compare their impact on the amount and focus of evidence obtained. Just as game designers can measure players' engagement and compare alternatives in play testing, the psychometricians can gauge evidence and compare alternatives for their relative contributions. Some of the points we will address involve classical test theory and others involve latent-variable psychometric models, such as structured item response theory, diagnostic classification, and Bayesian inference networks or Bayes nets.

A motivating goal of the GlassLab project under which our work takes place is to advance the practice of psychometrics for game-based assessment. The best way to do this is by integrating the principles of assessment design with game design, because applied psychometrics flows from assessment design. The patterns and parameters in psychometric models acquire practical meanings only through assessment arguments. It is thus assessment design that links psychometrics with game design. For example, we noted that game designers tailor familiar "game mechanics" schemas to situations and affordances for players to accomplish goals in a wide variety of specific situations in specific games. They combine principles of engagement and game play advancement, in re-usable ways. Similarly, assessment designers tailor task models and accompanying measurement-model elements to create specific tasks to elicit then manage information from examinees' performances about their capabilities.

We must therefore take some time to lay out key ideas and representations of assessment design, noting connections to games along the way and illustrating points with examples from Jackson City, to ground the discussion of issues that are psychometric in nature, but tap deeply and simultaneously into game and assessment design.

A concept we will keep in mind during the discussion and pull together in Chapter 13 is that of building blocks--very much like the idea of design patterns (e.g., Gamma, Helm, Johnson, & Vlissides, 1994), as clusters of generalized features of situations and potential actions that can be used / reused, which game designers have as building blocks especially for purposes such as advancing game play, increasing engagement, etc. This is important to us because they are analogous to task model and evidence model clusters, which are generalized features of situations and potential actions task which assessment designers have as building blocks to evoke and make sense of evidence in effective ways. In both cases, these are reusable clusters, or modules, because experience and theory have proved them to be useful to build specific artifacts (games, tasks) around as they optimize the design problem. We will be arguing later in the paper for the value of building block patterns that provide pre-packaged thinking about situation and affordance features jointly with regard to play and evidence i.e., GBA mechanics. See Mislevy, Steinberg, Breyer, Johnson, & Almond (2002) on how something like this was done to support the design of complex simulation-based problem-solving tasks.

The Running Example: Jackson City

GlassLab is a research and development effort funded by the Bill and Melinda Gates and the John D. and Catherine T. MacArthur foundations. The team is charged with creating digital games that engage students and measure learning. It brings together partners from the Institute of Play, Electronic Arts, Educational Testing Service, and Pearson to both re-engineer existing games such as SimCity that were designed for the broader market and to create novel game-based assessments.

GlassLab's first product is called SimCityEDU: Pollution Challenge! The game is a modified version of the current SimCity, a simulation that lets players plan, build, and "run" digital cities. SimCity, and other titles that followed such as SimEarth and SimAnts, emphasizes the agency and authorship of players, giving them a chance to create their own cities, ecosystems, and ant colonies, each populated with digital agents that mirror the decisions and activity of their real-life counterparts (Ito, 2009). SimCityEDU: Pollution Challenge! takes advantage of that perspective, providing students opportunities to build and create that are supported by an understanding of systems and human impact on the environment.

SimCityEDU: Pollution Challenge! is comprised of several missions in which students are presented with constrained, pre-designed, polluted digital cities. Successful completion of the missions requires students to plan and employ green technologies to reduce pollution while at the same time supporting their cities' economic growth. Working with the game's pre-designed cities, students are introduced to themes of human impact on the environment as presented in the Next Generation Science Standards (NGSS; Achieve, 2013) core disciplinary ideas and the NGSS' cross-cutting concept of systems and systems models (Table 1).

Jackson City is one of the more advanced missions in SimCityEDU: Pollution Challenge! During gameplay, students must find ways to 'decouple' economic growth from detrimental environmental impacts, growing their cities' economies while minimizing pollution. Figure 1 is a view of Jackson City a player sees at the beginning of a challenge. Players are first introduced to the challenge through a brief narrative and a request for help. And once they have accepted the challenge players then enter the three-dimensional city via a top down view.

While Jackson City is relatively contained, it presents the player with a rich set of manipulable objects. These include the major features you would expect in a real city - houses, large apartment buildings, rushing cars and buses, offices, power plants, industrial sites, roads, schools and parks, to name a few. Active sims – animated agents –move about the city carrying out their daily tasks – traveling to work, traveling to school, returning home, etc. By hovering their cursor over these individual sims, players can trigger a text-based script describing the sim's current experiences in the city. These casual and brief reports are determined by the current state of the sim's neighborhood, health and employment status. The sims are an important source of feedback to the player as they often voice their mood, reporting their views on the city's pollution levels, problems with unemployment, lack of steady electricity, the quality of the schools and levels of crime. The realism in the city's structures and residents is intended to increase the players' empathy for the virtual population and the city as a whole.

As a part of the challenge's introduction, the player is cast as the Jackson City Mayor and is also given access to policy tools that allow her to zone and rezone areas of the city, access to maps and gauges that reveal the quantities, concentrations and movement of pollution and help players spot patterns in unemployment. The set of action-objects also allow the player to bulldoze structures and build coal, wind and solar power plants, for example. The in game tools also support a high degree of authorship over the course of the city – allowing the player to decide the fate of residents, businesses and utilities for example.

The in-game tools are also necessary parts of a strategy meeting the Jackson City challenge to reduce air pollution while growing the economy and the number of available jobs – an optimization problem requiring players to plan ahead and work in a balanced way across several independent variables. First time players often start their gameplay by surveying the city, attending to the dark clouds of pollution and their origins, and then identifying the city's coal plants as the largest polluters. Once they have pulled up the bulldozing tool, these first time players typically bulldoze the coal-fired power plants, creating an energy shortage that is announced within the game and causes a series of ongoing brown-outs across the city. Ultimately, employment levels drop as a result and while the player may have met her pollution targets for the challenge such a strategy fails to meet the employment targets as factories and businesses close from the lack of power.

More sophisticated and successful strategies arise as players come to consider the multiple sources of the pollution (industries as well as coal-fired power plants) and appreciate the impact of each on the levels of employment within Jackson City. In that case, players typically come to view and use green sources of power such as solar plants and wind farms as helpful tools in lowering pollution while maintaining jobs. These players often build several green energy sources within the city early on in the game in order to build up their energy production and give themselves the chance to either turn off the coal plants or remove them completely. But even this level of coordination is insufficient to do well in the game, as successful players will also need to think about zoning policies that wean the city's job market from a reliance on pollution heavy industries in order to earn the top scores. In short, Jackson City poses an authentic and multivariate optimization problem, requiring players to minimize the city's impact on the environment while also maximizing jobs – attending to multiple variables in concert.

Assessment within the game depends on this multivariate nature of the Jackson City challenge to detect and measure students' facility with considering and intervening on systems comprised of multiple interdependent variables. Together, these are abilities that many in the learning and science education communities have come to call 'complex problem solving' or 'systems thinking' (Arndt, 2006).

Assessment within the game was designed to detect and measure students' facility with considering and intervening on systems comprised of multiple independent variables, and in some cases multiple dependent variables as well – altogether, capabilities that the learning and science education communities have come to call 'complex problem solving' or 'systems thinking' (Arndt, 2006).

Table 1:
Science Standards Addressed in Pollution City

Next Generation Science Standards: Cross Cutting Concepts

Systems and system models. Defining the system under study – specifying its boundaries and making explicit a model of that system – provides tools for understanding and testing ideas that are applicable throughout science and engineering.

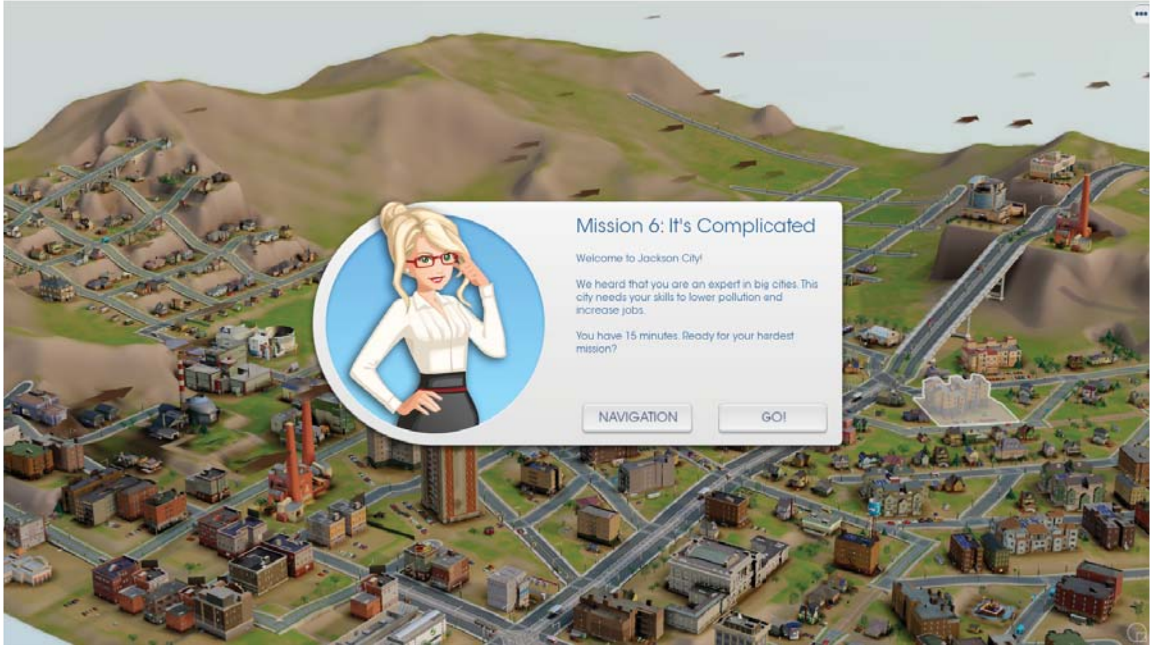
Next Generation Science Standards: Human Impacts on Earth Systems

Human activities have significantly altered the biosphere, sometimes damaging or destroying natural habitats and causing the extinction of other species. But changes to Earth's environments can have different impacts (negative and positive) for different living things.

Typically as human populations and per-capita consumption of natural resources increase, so do the negative impacts on Earth unless the activities and technologies involved are engineered otherwise.

Standards from Achieve, Inc., 2013.

Figure 1:
Initial View of Jackson City



Mission 6: It's Complicated

Welcome to Jackson City!

We heard that you are an expert in big cities. This city needs your skills to lower pollution and increase jobs.

You have 15 minutes. Ready for your hardest mission?

Where Do We Get Evidence?

Assessment is designing situations in which to obtain evidence about aspects of what students know and can do. Assessment can be done in many ways and for many purposes, from real-time assessment to guide learning in an ongoing activity, to external assessment of many students to provide information to policy-makers. By game-based assessment, we mean broadly an activity which is meant to obtain such evidence for some assessment purpose(s), and from the perspective of the student's experience, at least part of that activity has the feel and the features of a game. Our attention will center on games in technology-based environments, but this is not necessary for an activity to be a GBA.

It will be useful as we go along to distinguish what Jim Gee (2008) calls the "little g" and the "big G" senses of a game: The 'game' is the software in the box and all the elements of in-game design. The 'Game' is the social setting into which the game is placed, all the interactions that go on around the game" (p. 24).

The little g game, then, is the specific environment and activities of a game viewed strictly: what players are doing when they are said to be "playing the game"; chess players sitting at a table and moving pieces according to rules, for example. The Game of chess spans chess clubs, "white to win in three moves" puzzles in the newspaper, studying books on strategy, teaching a friend variations on an opening, and the excitement, the etiquette, and the mind games of tournaments.

This distinction is important because much of the learning and much of the socialization, that occurs in relation to games occurs outside the little g game. Gee argues that good game design is not just a matter of game design in the sense of in-game design, but Game design as well, or the design of the interactions around the game. These considerations as they apply to school learning lie at the heart of the Quest to Learn schools (Torres & Wolozin, 2011). Klopfer, Osterweil, and Salen (2009) give examples of how teachers learn to design the Game space that structures students' use of games. It is the exception rather than the rule that all of the important learning is expected to take place within the little g game. In the following section, we argue that the same broader view that is essential for good game-based learning is just as important for game-based assessment.

It is useful then to define three paradigms for where assessment and psychometrics can take place in GBA:

Paradigm 1: Assessment outside the game. One possibility is for all assessment to take place in the big-g Game but outside the small-g game. Here the small-g game is a location for exploration, play, learning, and problem-solving. Assessment would take place outside the game, based for example on students' solutions, their rationale as produced in a presentation, a write-up, or video they create in the game environment, and so on. These external-to-the-game work products would be evaluated, formally or informally, by automated or human means (e.g., evaluations by teachers, by external raters, or by students themselves with provided rubrics). An example is Digital Zoo (Svarovsky & Shaffer, 2007), where students learn engineering principles to design creatures that meet certain goals. The learning and exploration can take place within the game. The goal of the small-g game is to build creatures in the digital environment under various constraints, with various tools, that accomplish goals--such as walk.

Assessment design in such GBAs in this case would address targeted capabilities to be evidenced in students' working, features of their work, or qualities of explanations. Designers need to create game spaces, affordances, and challenges that evoke the targeted capabilities, and devise external-to-the-game work products to bring those capabilities out. The small-g game could be pre-existing (e.g., analysis of poker hands, either concurrent with play or following it), or created afresh. This kind of GBA is particularly well suited for capabilities that involve metacognitive and reflective capabilities. In fast-moving action games (as in sports) there can be value in analyzing what happened, why it happened, what it meant, and what to do. Connections, analyses, and deeper structures are more easily addressed in pauses, and allow for greater engagement during play. Psychometrics can be similar to those applied in performance assessments projects like the Advanced Placement Studio Art portfolio assessment. This approach requires minimal coordination among game code and psychometric/scoring code at the implementation level, although it would still require coordination at the design level to make sure the game activities address the targeted substantive content and the Game activities capabilities evoke the targeted knowledge and skill.

Paradigm 2: Assessment inside the game, prespecified work products. In this approach, as part of game play students must carry out certain actions or complete certain products that are pre-specified by designers. These prespecified products can be fairly simple, such as answers to arithmetic items in Math Blaster; very complex, such as a town plan with zoning recommendations in Urban Science (Bagley & Shaffer, 2009); or somewhere in between, such as planning forms, data forms that summarize results, or Punnett square representations to express hypotheses for animal breeding studies. They may be smoothly integrated into play, as they are in Urban Science, or noticeable and distinct, as when play stops as a student answers questions or evaluates the results of an interview with an avatar.

A design strategy that can serve game play and assessment jointly is for a pre-defined work product to be a logical part of the narrative, again illustrated by the reports to supervisors and final plans in Urban Science. These are examples of the joint game-assessment mechanics we will discuss later.

What is common among Paradigm 2 evidence production, however, are a set of defining characteristic features. They are ...

- Designed ahead of play, such that they
- Elicit evidence of targeted proficiencies in known ways, and
- Strategies for evaluating them, i.e., evidence identification routines, have been worked out ahead of time, at least provisionally.

Although they are part of game play, they are like familiar assessment in that they can be thought of as predefined "tasks" regardless of their complexity.

Paradigm 3: Assessment inside the game, evidence from work processes identified in data streams. (Shute, 2011). In more complex and interactive games, players have more choices about how to move through the game space, investigate situations, and meet goals. Students who all eventually reconfigure a malfunctioning router when they take a contract to troubleshoot a particular network in the Cisco Networking Academy's Aspire game might differ substantially as to how efficient and systematic they are, and whether they check the results of changes they make along the way. Thus features of sequences and partial solutions can provide evidence about their understanding of the network, their strategy usage, and their metacognitive skills along the way, over and above the evidence conveyed by their final solutions. Identifying and interpreting such data is one of the most exciting aspects of GBA, and one of the most interesting challenges to designers.

Assessment designs and psychometricians do not start from scratch in Paradigm 3, due to a rich, if small, tradition of performance assessment—that is, assessment where examinees carry out complex problem-solving or other pertinent challenges in real, hands-on environments or in simulations. The

National Board of Medical Examiners (NBME), for example, has been studying how to design and score computer-simulated patient-management case problems for more than thirty years, and has recently introduced them into medical licensure sequence in the form of simulation studies Primum (Dillon, Clyman, Clauser, & Margolis, 2002). Margolis and Clauser (2006) provide discussion on how the NBME designs these cases, identifies key features of candidates' widely varying solution paths, and provides final scores.

This presentation emphasizes Paradigms 2 and 3, or evidence of students' capabilities within the activities of the small-g game (although evidence outside the game as described as Paradigm 1 can be carried out as well). What students do and how they do it -- that is, product data and process data -- are potential sources of evidence. We will say more about kinds of data, and how they are identified and used, as we proceed.

Even within the category of formative assessment, the evidence can be used for purposes inside the game, outside the game, or both. The notion of feedback cycles is useful here, because modeling and action can take place in at different levels in hierarchies. In recreational games, for example, fine-grained counts and action monitors are used to adjust game situations moment to moment, while coarser status variables and play characteristics. There can be similar hierarchies for assessment in a GBA, such as fine-grained, local, modeling for feedback during play and coarser-grained modeling for a teacher's classroom dashboard. We will have more to say about kinds of data, and how they are identified and used, as we proceed. As we will continue to elaborate, we think of psychometrics in terms of managing information for what the various purposes in the various feedback loops in a particular GBA.

Use Cases: Roles for Game-Based Assessment

The word assessment covers a lot of territory. Assessments range from high-stakes certification tests and standardized college entrance examinations, to on-the-road drivers license tests and three hour long oral dissertation defenses, to a quiz in the classroom, a lesson in an intelligent tutoring system, and an informal conversations with a teacher. They can be meant to provide information to teachers, parents, researchers, intelligent tutoring systems, students themselves, chief state school officers, or potential employers or colleges.

To help frame the discussion, we can list some prominent roles that might be envisaged for game-based assessment. We borrow the term "use case" from software design to describe a configuration of actors, information, and processes that serve a recurring purpose. Seven are listed below, ordered from more intimate and immediate purposes to more external ones (note that the same GBA might serve more than one purpose). We speculate as to how well we think GBAs might serve these purposes.

- *Information for internal game purposes.* Gee (2007) argues that many good recreational games have already being doing assessment, at least implicitly. They gather information about aspects of a player's experience, success, what they are doing well and what they are not, in order to adjust the pace of play or the nature and level of challenges they offer—all to the end of providing an engaging experience. As a simple example, the system's matching the rival car's speed to a learner's current typing speed and accuracy is what makes the Car Racer game in Mavis Beacon fun. Some of the same deep principles in educational assessment are in play, although they don't, and don't need to be, formalized in psychometric theory or organized in terms of learning standards.

- *Formative assessment: Information for students.* A GBA can also provide information to a student as they play or at the end of sessions or challenges. Some information could be organized around details of what the player has done or accomplished so far, like the very detailed reports of weapons, energy, battle results, etc. of Civilization. Other information could be organized around standards as they apply to the specifics of the challenge, or in terms of progress with respect to standards or learning objectives as they relate to progress in the game challenge. The latter kind of information could draw more profitably from educational assessment methodology to manage evidence and uncertainty, as well as from the literature on formative assessment (Black & Wiliam, 1998; Heritage, 2010). These latter kinds of reports can be useful in calling students' attention to higher-level or cross-cutting ideas, promoting reflection beyond the immediacy of actions within the flow of play.

- *Formative assessment: Information for teachers.* Teachers working with groups of students could also use information of the second type from the previous use case: Summaries of how students are coming along with respect to challenges and learning objectives, so as to keep a class on pace, lead classroom discussion on key concepts, or trigger follow-up with certain students. Something like a "teacher dashboard" could be useful, again drawing on educational and measurement experience. Discussions of how a challenge fits in with a cross-cutting idea such as energy transfer that appears in different guises in different areas may be better facilitated by these conversations than by mechanics situated within game play.

- *Information for designers.* If many students are playing a game, information about aspects of play such as feature usage, heightened or decreased engagement, sticking points, and pacing can be gathered and explored to improve play and improve learning (El-Nasr, Drachen, & Canossa, 2013). This kind of information is routinely used by game designers to improve play, and more recently by instructional designers to improve learning (e.g., Koedinger, Aleven, & Heffernan, 2003). In similar ways, GBA design teams can use the information to improve the balance among game and assessment objectives.

- *End of course assessment.* End of course assessments might be able to use segments of games to evaluate learning in a course, if sufficient groundwork has been laid: Students have become familiar during the course with the topics, representations, interfaces, and expectations of the game. It can be possible to carry out moderately high-stakes assessment in this case because these sources of construct-irrelevant variance among student performance have been mitigated, and what will make the game challenging is the learning objectives, not the game per se. For use at this level of stakes—a course grade, for example—it is appropriate to use methods for addressing reliability and validity more formally.

- *Large-scale accountability assessment.* A topic of much current interest in education and assessment is instruction and assessment based on subject-area standards, such as the Common Core State Standards in mathematics and English Language Arts (Common Core State Standards Initiative, 2010a, 2010b) and the Next Generation Science Standards (NGSS; NGSS Lead States, 2013). Large-scale accountability tests are planned at the state level, in which all students in a given grade would be administered an assessment based on standards for their grade levels. Stakes for students, teachers, and/or schools might be attached to the results. Given the potential of GBA to increase engagement, game-based tasks might be contemplated for use in such assessments. Engagement requires investment, however, and the same deep features that draw some students into a task and provide better motivation can be unappealing to other students and provide less information about their capabilities. Issues of familiarity with interfaces and expectations, and wide variation across task content (i.e., "low generalizability"; Linn, 1994) and narrative features also militate against using GBA in settings that are both high-stakes and "drop in from the sky" (that is, they have no direct relationship to what students are studying).

- *Large-scale educational surveys.* Educational surveys such as the National Assessment for Educational Progress (NAEP; Jones & Olkin, 2004) present samples of tasks to samples of students in order to provide a snapshot of what students in a state or country are able to do. These assessments drop in from the sky, but they hold no stakes for individual students, teachers, or schools. Some use of GBA could be justified in these assessments, to provide information to researchers and educators about students' capabilities in such environments and to learn more about the variation that argues against their use in high-stakes use case described above.

A Sociocognitive Perspective on Learning

Game design, instructional design, simulation design, and assessment design each have their own goals and methods, but to design artifacts in the intersection it helps to have a common psychological perspective on which all are necessarily grounded. We take a sociocognitive or situative perspective (Gee, 1992; Greeno, 1998; Lave & Wenger, 1991).

The "socio-" in "sociocognitive" highlights the patterns of knowledge and activity that structure the interactions people have with the world and other people. These include the structures and ways of using language, knowledge representations, and cultural models, and of the patterns of activities of families, communities, personal interactions, and classrooms and workplaces. Of particular interest for our purposes are the kinds of things we learn for school and work: skills, knowledge, identities, values, and epistemologies (SKIVE elements, as Shaffer, 2007, calls them) for working with scientific models, for example, or troubleshooting computer networks, or developing zoning plans for communities.

The thing about game-based assessment that makes them interesting and also makes them hard to design is that they incorporate semiotic patterns from multiple domains simultaneously. A student playing Jackson City draws on linguistic, cultural, and substantive patterns of many kinds and at many grainsizes. She must understand something about mayors, cities, jobs, and power plants; maybe not zoning, but enough about the others to learn quickly. She must understand English well enough to make sense of help, scenario descriptions, and simulated citizens' complaints. She must navigate in a SimCity style world, moving from one view to another, and do things like zoom, plop, and hover. She must coordinate her play and understanding of Jackson City with all of the activity patterns and knowledge patterns of the classroom, particularly the ones that create the big-G game that envelop her actions in Jackson City. And the whole point is that even given all this knowledge, she may not know much about how elements of systems act together and how to talk and think about systems—and interacting in this artificial world, learn something about how to talk about and think about systems more generally.

The "-cognitive" in "sociocognitive" highlights within-person cognitive patterns, from large to small and across different levels—all traces of each individual's past experiences, continually assembled, adapted, and revised to make meanings and guide actions in each new situation. A sociocognitive psychological perspective addresses the interplay among these levels: Neurological processes within individuals give rise to their actions in the human-level activities we experience, as we negotiate the physical and social world.

The key to developing capabilities in some area is acting in situations that in some way develop familiarity with recognizing what is important in that area, working with the representations and the language, learning the ways to act and think, getting feedback from other people or the situations themselves. Learning to troubleshoot networks, for example, might involve listening to lectures and reading texts, to build up certain knowledge structures, but becoming proficient will inevitably also require identifying faults in real networks or simulated ones, usually with support from others (Lave & Wenger, 1991). The goal in education is helping students develop resources for recognizing, thinking about, acting in, and creating situations, through the knowledge structures and activity structures of the domain. The goal is internal, but the situations for learning in are external. Just how to structure situations and activities depends on a designer's purposes:

- Instructional design is determining situations for students to act in to develop resources. What are key features, what are the goals for students, what affordances should they have, how might activities be sequenced and supported, what are good mixes of different types of activities?

- Assessment design is determining situations for students to act in that give clues about what they know, how they are thinking, what how they interact with problems, and so on, to provide information as feedback on the learning. It might be information for an instructional system, the students themselves, a teacher, a researcher, or a chief state school officer, in each case with their own purposes and contextual knowledge.

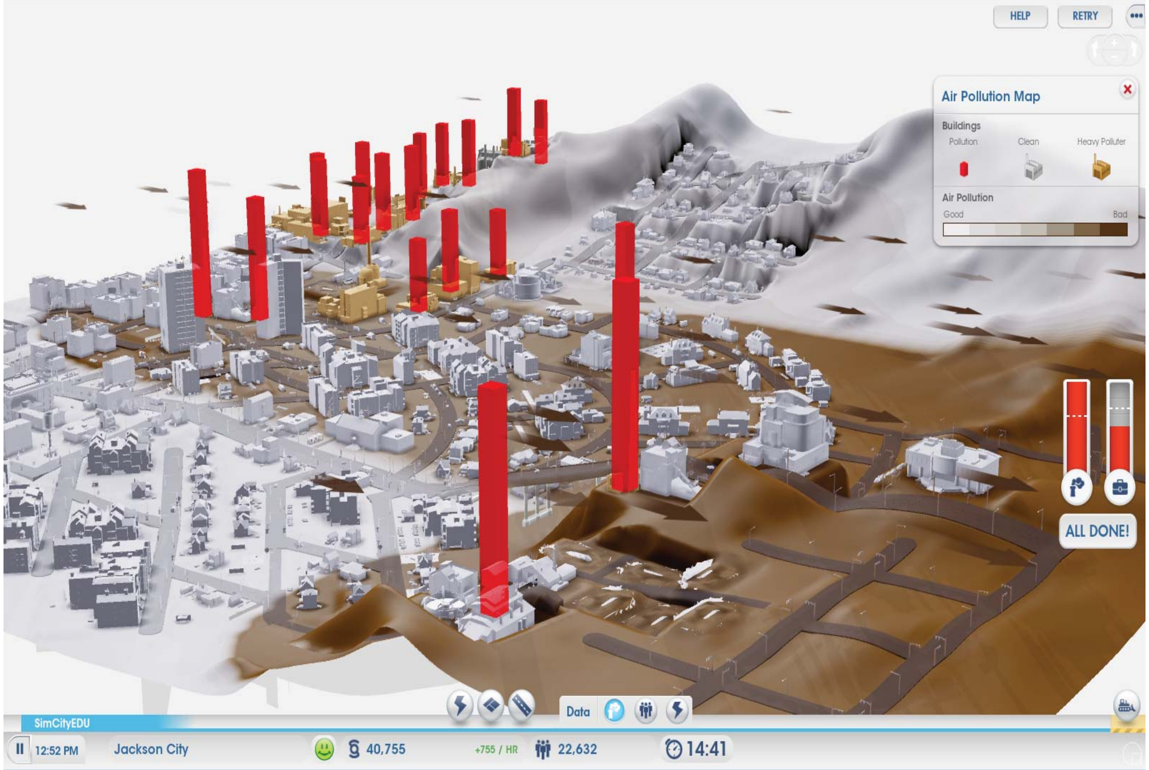
- Game design is determining situations that engage players – which turn out to be situations at the cusp of their limits, so the objective is figuring out goals, situations, story lines, and affordances to adapt play to keep them in that neighborhood (Gee, 2007). Learning, engagement, and information all tend to be high when people work near their frontiers, so game design, assessment design, and instructional design work together on this point rather than compete.

- Simulation design, in the contexts of instruction / games / assessment, is identifying those features and affordances of situations to incorporate in a simulation environment, which to enhance, how to represent them, and which to ignore, so that activity will be most edifying / engaging / informative (Roschelle, 1997).

Instructional design and game design work jointly in Jackson City to support students' comprehending and intervening in the complex systems in the game. The game's missions reflect increasing levels of complexity of systems thinking. They are based on a progress variable for systems thinking derived from earlier work (Brown, 2011; Shute, 2007), describing how students are likely to progress in their facility with complex systems. The progress variable's levels range from more naïve understandings of the given system where players have little or no awareness of the independent variables involved, to more sophisticated cases in which players are aware of and manipulating multiple independent variables in order to direct change in one or more dependent variables. The missions reflect the progress variable and become increasingly complex as students proceed through the game.

More personalized player-supports have been put in place as well. Figure 2 shows a view that players can use to monitor sources of pollution. If players have not accessed key maps, indicators or tools within specified timeframes during game-play for instance, dialogue pop-ups are presented to the player to draw their attention to particular features of the game and describe the role of those features in meeting the mission's challenge. Such in-game feedback is one of three levels of feedback that have been built into the game: in-game feedback, mission feedback, and summary-level feedback. Where in-game feedback is meant to help direct the student to a successful performance without impacting their understanding of the system, supports offered through the mission feedback and the summary-level feedback are designed to help students identify their likely position on the systems-thinking progress variable and motivate reflection on how they may improve their performance.

Figure 2:
Use of a Tool to Monitor Amounts and Locations of Pollution Production



Elements of Game Design

There are many ways to define games (Salen & Zimmerman, 2004). Games can be viewed from a cultural perspective as expressions and representations of historical events, symbols and rules (e.g., *Civilization*), a learning perspective as representations of real world phenomena in a restricted environment (e.g., flight simulators), a societal perspective as mechanism to compete and rank (e.g., *Madden NFL*), and numerous other perspectives. Yet, across those perspectives, most games contain several critical key elements that each, or in combination, result in effectively establishing an intuitively interactive system that provides engaging experiences. In discussing game elements, we distinguish two types of interrelated elements: building blocks that comprise the architecture of the game, and elements that characterize the experience of the game.

Game development iterates between these two types of elements, influencing and representing each other in many ways. The fundamental challenge for game based assessments is to find the common ground where both architectural and experiential game elements either coincide, enhance, or, at the very least, do not undermine key assessment elements (e.g., evidence gathering, data retrieval, design patterns that indicate how to obtain evidence about targeted capabilities), and vice versa.

Architectural game elements

The elementary particles of a game are objects, rules, connections, and states. Rules define what the connections between objects are (i.e., how they behave and interact). Together they provide an account of the current state of the game and changes to that state. Sets of particles can form higher-order game play elements.

Objects and Rules

The basic structure of most games revolves around rules that define how the game reacts to player behavior (including inaction), given the current state of the game. In its most basic form, a rule is a function:

$$y_{(t+1)}=f(x_t\,|a_t,b_t,c_t)$$

where $y_{(t+1)}$ is the reaction of the game at time t+1, which is a function of the user behavior x at time t, given conditions at time t of features a_t, b_t, and c_t, that reflect the current state of the game at time t. Subsequently, a,b and c are updated to reflect the fact that $y_{(t+1)}$ happened.

Most games make use of objects with attributes (e.g., a two or three dimensional mesh frame of a tree that has leaves and bark artwork associated with it for display purposes), which can also be defined as a set of rules for properties that apply in particular circumstances (e.g., rules about when the physics of wood apply). The environment itself is an object or a set of rules. Many objects are reusable (e.g., a forest is created by reusing tree-objects that are placed at slightly different angles and with different shades and colors), and can be anywhere from highly complicated (e.g., a simulated person that can interact in thousands of ways) to very simple (e.g., a background picture to give the illusion of a scenery).

This general, functional definition of rules and objects connects in certain ways to assessment elements. Assessment tasks can be viewed as objects and rules, which can be fairly simple in multiple-choice based assessments and very complex in simulation and game based assessments. Stimulus materials and response options in multiple-choice tasks are rather simple objects, and a rule associated with an object can be to darken a radio button and send a message containing the choice to a response-evaluation process. In a GBA, the interactions of a player with objects can not only provide evidence for inferring a player's strategies or proficiencies, they can change the situation in ways that provide at once further game play and set the stage for acquiring more evidence. A GBA can incorporate objects that play a key role in evoking and gathering evidence and simultaneously serve a role in game play that may be seamlessly integrated, incidental, or (disconcertingly) disconnected from play.

Connections

Connections make explicit the relationships among all the objects and their attributes, and therefore indicate how the rules work synchronously. Rules are particular kinds of connections. Connections are the building blocks for experiential game elements such as narrative, goals, feedback, and rewards. Patterns of actions under certain game states are also the source of evidence for the assessment aspect of a GBA; e.g., sequences of actions in states with particular features (e.g., in Jackson City, building a replacement green power plant before bulldozing a coal plant), or attributes of an object (in Aspire, do security settings block and allow the desired messages to the PC in the student lounge?).

GBAs have distinct networks of connections—one that makes elements and actions into a functioning game, from the player's point of view, and another that makes elements and actions into functioning assessment, from the point of view of the user(s)—the player, the system, and/or the teacher. It is the networks of connections that make objects and actions function as games and assessments. Objects and actions that are important to gameplay and those that are important to assessment can be distinct and obvious to the player, or they can overlap more substantially and feel more seamless—even to the point that the assessment functioning is unnoticeable (e.g., Shute 2011).

State

The state of the game defines what actions are possible at a particular point in time. In most games, the state can also provide some information about the history of states. For example, in Diablo, being in the state "in the dungeon four levels below the surface" means the player has mastered the first three levels. After player's actions, rules update the state. When Tomb Raider Laura Croft enters a plain room, she may be able to run, walk, jump, crouch, shoot, or push buttons, whereas when she enters a pool, she can only swim, dive, or push buttons.

The state of a game is often connected to the level of a player at a particular point in the game. A more advanced player in Everquest has bigger swords to fight bigger monsters and to earn more points to buy even bigger swords. Here the state of the game translates directly to the proficiency level of an examinee at a point in game play, and as in adaptive testing, the difficulty of a challenge ('monsters') is just above the estimated competency ('sword size') of the student, to produce an engaging experience.

A vector of game condition variables (just a few in a simple game, thousands in more complicated games), in conjunction with rules that govern possible actions, interactions, and behaviors of objects is called a state machine. We will see that the state machine in a GBA can be extended in a natural way to assessment functioning as well, in ways that can be understand through a four-process architecture for assessment interactions (the Assessment Delivery section in Chapter 6).

States in Jackson City game include the number of city objects such as homes, roads, factories, cars and different types of power plants, among others. The states also include variables such as the amount of air pollution, the location of the pollution and the direction it is traveling, the number of available jobs and the number of students in a given neighborhood that are enrolled in school – among others. As students intervene on the various objects within the game – bulldoze power plants, zone for new houses, etc. – students act on and change the game's state. In the challenges on systems thinking, students whose understandings of the system are more sophisticated are likely to have very different end-states for their cities than students whose understandings are less sophisticated. For example, more sophisticated students will tend to create cities with fewer coal power plants, more commercial jobs, and fewer industrial jobs as they work across multiple independent variables driving pollution while also working to maintain jobs. Less sophisticated players' cities may have low amounts of pollution but they may leave untouched the proportion of city-dwellers who are employed in commercial spaces and those working in industry. These contrasting states begin to provide some evidence for distinguishing between the players with regard to their ability to intervene effectively, given the systematic relationships that underlie their city's economy and ecology.

Mechanics

The term "game mechanic" combines elements discussed above, to describe configurations of kinds of actions that players can take in recurring situations in a game, with certain kinds of outcomes on the game state, to advance play. The video game Angry Birds uses the mechanic of sling-shotting birds. The Gamification Wiki has a taxonomy of common mechanics in video games (http://gamification.org/wiki/Game_Mechanics).

A game mechanic is an engineering concept, but using a mechanic is central to the players' experience. It is through the mechanics that they experience a game: How they recognize what is important, what it means, what they can do, what might happen if they do it, and how sequences of actions through these mechanics can combine into tactics and strategies

In GBAs, a designer wants the kind of thinking that game mechanics evoke to advance play to also promote the targeted thinking in the domain. Ideally, how the player learns to act and think in order to do well in the game overlaps significantly to how one must act and think in the target domain (Shaffer, 2007). The mechanics are designed to structure the player's thinking in this way, and their actions through the mechanic both advance play and provide evidence of their thinking (Plass, Homer,Kinzer, Frye, & Perlin, 2011).

In Aspire, for example, players configure, replace, and connect network devices. These objects, with their configurations being attributes, and their rules determine how they send data (or don't) in accordance with their built-in rules and current configuration values. To execute a maintenance contract, a player uses mechanics that are virtually the same as the ones actual network engineers use to troubleshoot actual computer networks

While *SimCity*™ itself may be better identified as a *construction game* (Ito, 2009) several game mechanics are at work in *Jackson City*. For example, they can bulldoze, dezone and rezone areas of the city, and create ("plop") wind and solar power plants and other kinds of buildings. This connects with assessment in that their choices and sequences of actions of these kinds gives clues about first their exploration of how the system components interact, and then their manipulation of system elements to achieve goals as it functions over time.

Experiential game elements

A game designer determines the kinds of situations players will encounter, how they can interact with them, and what they want to accomplish. Engagement depends in part on their perception of autonomy, competence, and relatedness. Autonomy refers the extent to which a player is given control over and freedom to choose his or her actions. Competence refers to the extent to which a player can gain, demonstrate, and apply skills. Game design concepts here include goals and challenges, complexity and discovery, feedback, and adaptation. Relatedness refers to the extent to which a player can identify, collaborate, and foster empathy. Game concepts are compelling narrative or setting, and social aspects of games. Competence is closely related to learning and assessment.

Game features that increase autonomy and relatedness can increase engagement. However, from an assessment perspective, increased autonomy can decrease the comparability of evidence across players, and features that increase relatedness can introduce construct irrelevant variance.

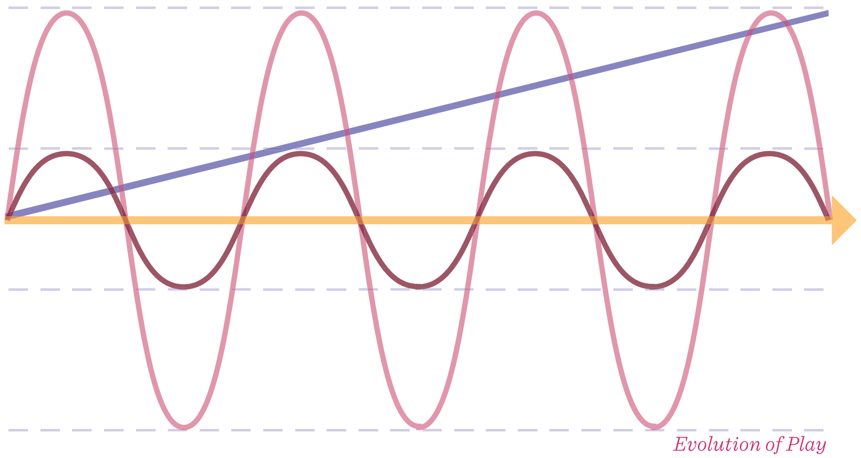
Engagement also depends in part on creating situations, challenges, rules, and affordances that will keep players near the leading edge of what they can do (Gee, 2007). This is one finding where game design, instructional design, and assessment design roughly agree: It is around the cusp of their capabilities that people experience what Csíkszentmihályi (1975) called "flow," what Vygotsky (1978) called the zone of proximal development in learning, and Lord (1970) called "maximum information" in computerized adaptive testing.

Autonomy

Autonomy in the context of games and assessments can be viewed as the ability to make choices about which experiences to engage in next and whether to continue in a particular experience or change.

Authorship. Authorship concerns the constraints that are placed on a player's interaction with the game. The most restrictive case is akin to a movie, where the viewer has no influence on how the story develops. The straight line in Figure 3 suggests there is no variation in the situations experienced by different participants. The least restrictive case is where only an environment is provided and players create their own game (e.g., Second Life). The large waves in Figure 3 suggest a great deal of variation in players' experience as they determine their own story lines and goals. Most successful games find a middle ground, where there is a clear, compelling story line, but also where players are allowed to deviate substantially before they are pulled back. These solutions provide the reward and enjoyment of discovering and learning about an existing story or environment, but also allow for substantial autonomy to make important choices about how to solve tasks, effect individual preferences for how to interact with the game, and in some cases even to contribute design elements that can be distributed and used in a larger user community. Providing choices about how to solve tasks and in some cases creating design elements to foster autonomy will also be important to assessment functions in GBA as well, for providing evidence about what students know and can do.

Figure 3:
Degree of Authorship Afforded to the Audience in an Experience and its Affect on Narrative Direction



🟨 Small Player-Determined Deviations Around the Main Theme

No authorship by the player

Total authorship by the designer

🟫 Small Player-Determined Deviations Around the Main Theme

Some authorship by the player

Designer authors main theme and permits some player deviation before pulling them back

🟥 Large Player-Determined Deviations Around the Main Theme

High degree of authorship by the player

Designer authors main theme and permits player deviation and provides infrequent or optional plot points

🟪 Open-Ended Experience

Total authorship by the player

Designer creates the environment or provides a premise

Chaos, Determinism, and Opacity. Games often rely on a certain level of chaos, to both simulate real-life, unpredictable variation and to reduce the number of deterministic actors and events that need to be developed. Deterministic events are only as powerful or interesting as they are complex, which is a challenging proposition to author. While some level of chaos can increase the realism of the narrative, more substantial levels reduce meaningful autonomy for making informed choices and result in frustrating game play. Similarly, opacity indicates the extent to which the underlying mechanics of the game are made known to the user – in the sense, that is, of the nature and behavior of the objects and the rules as they are experienced in play. Complete lack of opacity, or transparency, does not allow for much discovery, a critical element in formative assessment. Too much opacity makes a game impenetrable and is experienced by the user as chaos, removing all sense of autonomy.

The designers of Jackson City want enough predictability in the interactions of the jobs and pollution factors for players to be able to discover their mutual influences by changing and comparing different states of the city they can create, but enough unpredictability that trying out arrangements gives only probabilistic and evolving evidence, as the simulated residents of the city carry out their actions under the new configurations players create.

Competence

Being able to apply competence and to demonstrate competence are essential elements for building self-esteem and intrinsic motivation in games.

Goals and Challenges. A central component of each game is the goal of the game: What is the player supposed to achieve? The goal is often how a game is first or predominantly described. It is a deciding factor for a player to play the game, as he or she assesses what kind of competencies are likely needed and whether he or she has the skills, or can obtain through the game, to succeed. Successful games provide the opportunity to develop a new competency that is at once challenging and achievable. This is also one of the underlying mechanisms of learning as a strong, intrinsic reward that signifies substantial overlap across games, learning, and formative assessment. Challenges are intermediate opportunities for discovery and mastery of new (sub)skills.

A Jackson City player's goal is to reduce pollution without putting citizens out of work. Doing so will require actions that jibe with the underlying dynamic system. How the player tackles the problem and how well she succeeds will provide evidence about her (improving, we hope) understanding of this system.

Complexity and Discovery. Complexity indicates the level of skill and understanding of the objects, rules, and game state a player needs to meaningfully engage with the game. Complexity is the experiential counterpart of a game's opacity. A designer wants to present a progression in complexity as players gain skills and can discover and master more complex situations, while avoiding an overwhelming level of complexity that quickly leads to disengagement. Metacognition research suggests a basic progression of (1) not knowing what you don't know, (2) knowing what you don't know, (3) knowing something but not realizing it yet, and (4) realizing what you know applies. An engaging game that provides ample opportunity to gain and demonstrate competences cycles players from (1) to (4) quickly and often.

Feedback. For most games, feedback is direct: defeating an enemy and continuing to live, or performing a sequence and being allowed to move to a previously closed place. Feedback in the form of rewards such as points or diamonds is part of many games. These tokens can be used to buy upgrades, unlock levels, or rank on a leader board. Direct feedback is usually limited to a particular task, while tokens can be viewed as indicators of a broader underlying competency that can be quantified across tasks, similar to a score in an assessment. The process of assigning scores and evaluating evidence might differ in games and assessments, but feedback is a critical component for players to determine levels of competency and to develop heuristics for increasing competency, therefore, creating autonomy.

Adaptation. Games that are relatively adaptive towards the level of the player and are willing to relinquish some authorship will likely be able to provide a richer learning experience. Both games and assessments have a rich history of adaptation, although there may be some conflicting objectives or uses. A basic goal for a game is for players to play it many times and for long durations. That is, extending the shelf-life (i.e., the time it takes before a game lands permanently on the shelf) is a measure of success. One way to accomplish this is to provide enough adaptivity in order for every player to, ultimately, win. On the other hand, Holland (1994) notes that many educational tests can be viewed as contests, where the purpose is not for everyone to win, but for only few to win access to a scarce supply such as scholarships or admittance to Ivy league schools, among a plentiful demand, in that case high school graduates. Adaptivity in the form of computerized adaptive testing is designed to reveal differences most efficiently, a very different use of adaptation. The contestant view is grounded in a summative assessment framework while and the game-like adaptivity that 'tries to help everybody win' is more consonant with the formative assessment purposes of learning games like Jackson City.

Relatedness

Recall that relatedness refers to the extent to which a player can identify, collaborate, and foster empathy in a game. While relatedness has not shown to be as strongly associated with intrinsic motivation compared to competence and autonomy, it does address the importance of a compelling context for an engaging experience and a social, collaborative nature.

An important component for simulation-based and role-playing games is a compelling narrative or setting that relates in a significant way to the (mental) world(s) of the player and creates attachment. Essentials of good storytelling are present, such as character development or logical progression of the story, and elements such as suspense, mystery, controversy, temporary hardship, humor, plots, themes, and adversaries are present and plentiful.

The challenge in recreational games is to present a narrative that is compelling to a wide audience. For assessments, a compelling narrative and well defined context is equally important in performance-type tasks, and can increase engagement and remove some sources of construct irrelevant variance as it clarifies the objective to the examinee. An example is the NBME Primum patient management cases in medical licensure. This contextualization in a narrative also presents a challenge in terms of assessment, however. The performance can become highly context dependent, a different potential source of construct-irrelevant variance, as the usual goal is to make inferences that hold across contexts at a more general level. For example, an improved understanding of gravity in a game where different size vehicles need to be driven on the hills of different sized planets (e.g., 'Hill Climb' by Fingersoft) may not transfer to a game where water needs to be directed into an alligator's shower pipe (e.g., 'Where Is My Water' by Disney), despite the use of the same underlying scientific principles.

Social Games. The Massive Multiplayer Online Role Playing Game (MMORPG) genre has gained immense popularity, in no small part because it directly satisfies the need for relatedness. Everquest and World of Warcraft are examples of MMORPGs that provide a great deal of relatedness as players can form collectives to achieve the goals of the game more quickly, but also to simply experience the game collaboratively, for example, making unique contributions as a particular kind of member of a team. In contrast, assessments, particularly summative assessments, are generally geared towards assessing individual competencies in order to make comparisons for the purpose of evaluating individuals or distributing a scarce resource fairly. Social aspects of assessments may occur before an actual test, such as studying together, but social interaction during a "test as contest" is usually cheating. However, as attention focuses on 21st Century skills such as collaborative problem solving, an honorable sense of social assessment may take hold. We will say a bit more about modeling collaboration in Chapter 11.

Assessment Design

We are applying a framework called evidence-centered assessment design (ECD) to the assessment-design aspect of game-based assessment. ECD views assessment as a special case of evidentiary reasoning -- that is, reasoning from a collection of particular nuggets of data obtained under particular circumstances, to broader interpretations of what students know, what they can do, or how they are thinking, or to actions based on interpretations like these. Fuller accounts of ECD can be found in Almond, Steinberg, and Mislevy (2002), Mislevy and Riconscente (2006), and Mislevy, Steinberg, and Almond (2003). More focused treatments of ECD and related psychometrics for game-based and simulation-based assessments appear in Levy (2012), Mislevy (2013), and Shute, Ventura, Bauer, & Zapata-Rivera (2009). The following sections summarize the key ideas that are necessary to an integration of assessment design with game design, thus laying the foundation for psychometrics for GBA.

Arguments and Layers

Two overarching ideas organize this sketch of ECD: arguments and layers. The first idea is seeing assessment as an argument from limited evidence. Messick (1994) says:

> "We would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors?" (p. 16)

This basic narrative takes any variety of forms for different kinds of assessment. As we go along, we will note how it plays out in familiar multiple-choice standardized tests to fix ideas, then see how it extends to less familiar forms like simulation-based and game-based assessments.

The second idea is distinguishing layers at which different kinds of activities and structures appear in the design and implementation of assessment, all to the end of instantiating an assessment argument in operational processes (Mislevy, Steinberg, & Almond, 2002; Mislevy & Riconscente, 2006). The layers shown in Figure 4 focus in turn on the substantive domain; the assessment argument; the structure of assessment elements such as tasks, rubrics, and psychometric models; the implementation of these elements; and the way they function in an operational assessment (for us, a GBA). The layers are distinguished by the kind of work that takes place in them, rather than representing a waterfall work flow process—that is, starting from Domain Analysis and working through each layer in sequence without cycling back. There are certain natural workflow implications: You can't implement a prototype without having done some thinking about what's important in the domain, for example. But in practice, we generally see moving cycling and refinement, an iterative design process learning and detailing as it progresses. Chapter 13 will describe a design process that synthesizes the work in ECD layers with the agile design philosophy typically used in game design.

Figure 4:
ECD Layers

Domain Analysis
Standards, Research, Previous Assessment, etc.

Domain Modeling
Assessment Arguments, Design Patterns

Conceptual Assessment Framework
Student, Evidence, Task Models

Assessement Implementation
Detailing Scoring Routines, Fitting Models, Authoring Tasks, etc.

Assessment Delivery
Four-Process Delivery System

Domain Analysis: What is important in the domain?

The Domain Analysis layer is concerned with gathering substantive information about the domain of interest that will have implications for assessment. This includes the content, concepts, terminology, tools, and representational forms that people work with in the domain. Equally important are the situations that people use that knowledge, and the things they do. For learning, this domain research provides us with information about the kinds of situations, representations, goals, and actions we need to build into an environment so students can learn. For game design, it tells us something about goals, narratives, and activity structures within which players will work. For assessment, it tells us something about how we will need to craft situations so that players' actions will give us clues about their understandings and capabilities.

As an example, Jackson City is addressing the Next Generation Science Standards' cross cutting concept "Systems and Systems Modeling" (Table 1). Students are expected to demonstrate an understanding of how to delineate the components and boundaries of systems, as well as how to represent systems in order to understand and test ideas or claims about them.

The standards are not sufficient in and of themselves to design and interpret assessments, let alone game-based assessments. They say little about the ways that students develop these proficiencies, or the situations and activities that both foster them and provide starting points for assessment. We therefore draw on research in the science education literature to help us understand students' scientific reasoning capabilities and how they develop, such as Goldstone and Wilensky's (2008) article "Promoting transfer by grounding complex systems principles," Sadler, Barab, and Scott's (2007) "What do students gain by engaging in socioscientific inquiry?", Brown's (2005) "The multidimensional measure of conceptual complexity," and Cheng, Ructtinger, Fujii, and Mislevy's (2010) "Assessing systems thinking and complexity in science."

Domain Modeling: The Structure of Assessment Arguments

Domain modeling is about how one might arrange features of assessment situations (everything from multiple-choice tests to GBAs) so they evoke the targeted knowledge and skills, and have students say or do something that provides evidence about them. The ideas, the representations, and the resulting design discussions are meant to be accessible to all members of a design team. For a GBA, this includes game designers, psychometricians, subject-matter experts, teachers, psychologists, and anyone else whose knowledge needs to come together to design the GBA. Domain Modeling is a work space where these people can share ideas and sketch out ways that the situations and action in the GBA might play out. This is where they recognize and begin to balance considerations from all their areas jointly. Successive approximations that begin to address competing constraints are cheaper to start early, in contrast to recognizing conflicts only after much time and money have been spent. Experts from each area will each have their own language and tools from their own area that they will need to bring to bear in the GBA (e.g., game designers' mechanics and psychometricians' measurement models), but it is in discussions at the domain modeling layer that goals and ways of attaining them, and constraints and ways of satisfying them, can be discussed across specialties.

The Messick quote cited earlier is a good start for understanding assessment arguments, but we need to elaborate it to design assessment tasks and GBA situations. We can build on philosopher Stephen Toulmin's general schema for arguments, shown in Figure 5.

In assessment, the claim refers to the target(s) of inference in the assessment. It might be some educational competency such as level of proficiency in scientific problem-solving, or something much narrower such as whether a student is systematic or floundering in a particular troubleshooting phase. Claims, and the data to support them, depend on needs for information in feedback loops.

A GBA can encompass multiple feedback loops at different grainsizes for different purposes. We propose data—such as quality of responses to questions or justifications students give for hypotheses—that support the claims. The *warrant* is the rationale for why certain observations might be useful evidence for certain claims. *Alternative explanations* are especially important in assessment arguments because they are central to validity. We may want to make inferences about students' competence based on their actions in a game, but are there other ways they could have done well without understanding, say, transmission mechanisms? Might they have struggled not because of modeling skills but because of time pressure? Is a player distracted by extraneous features of the game?

Figure 5:
Toulmin's (1958) Structure For Arguments

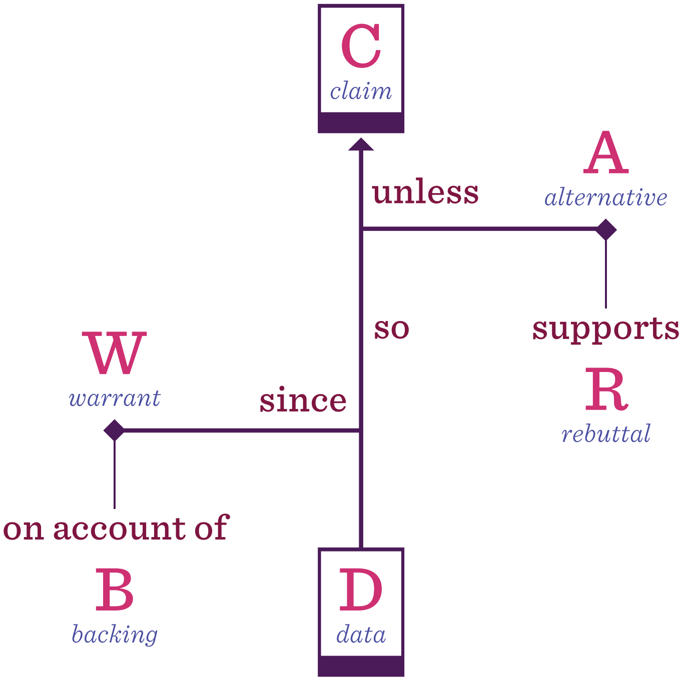


Figure 5
Adapted from Figure 1 from Mislevy, R.J. (2005). Issues of structure and issues of scale in assessment from a situative/sociocultural perspective. CSE Technical Report 668. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Copyright 2003 by The Regents of the University of California. All rights reserved.

Reasoning flows from data (D) to claim (C) by justification of a warrant (W), which in turn is supported by backing (B). The inference may need to be qualified by alternative explanations (A), which may have rebuttal evidence (R) to support them.

Figure 6 adds detail for the particular kinds of argument we need to make when we design assessments and interpret results. To illustrate claims, we will use the running example from Jackson City shown in Table 2. It is a learning progression for systems thinking, based on the research on systems thinking (Cheng et al., 2010) and increasing sophistication in scientific reasoning (Brown, 2005) mentioned in the section on Domain Analysis. Observing what players do, we will want to make inferences about their level of reasoning in the GBA more generally in these terms. [2]

[2] We do not expect a description of a level to characterize a given student universally across systems and contexts. Evidence suggests that peoples' understanding of systems can vary substantially from one system to another; that increasing understanding need not follow well-defined levels; and different situations can evoke thinking at different levels even within the same person (Sikorski & Hammer, 2010). Rather, we use the learning progression to manage situations and demands in the game, and to organize a probabilistic summary of patterns of "noisy" performance of students as they work through challenges with increasingly complex aspects of systems. We can use the learning progression to help design situations and manage evidence, without having to take it as a "faithful" model of students' capabilities.

Note that the assessment argument (Figure 6) depicts three distinct kinds of data. The first is features of students' actions, which is what people are familiar with as "data" in assessment. In Jackson City, for example, players can bulldoze existing power plants and build new ones that run on different fuel.

But just as important, from the sociocultural perspective, are the features of the situation the student is acting in. These are the second class of data. What we call Observable Variables (OVs) in assessment generally involve both: performance in certain features in a situation with certain features. Key features of situations are typically designed in to traditional assessment tasks and presented to examinees pre-constructed. There can be some identifiable and preconstructed "tasks" in simulations and GBAs, but we can also seek patterns of performance in situations that are unique to examinees as they work through a less constrained environment. We will say much more about this in Chapters 8 and 9.

A Learning Progression for Systems Thinking

Level	Competency level Description
1	Students have a fragmented understanding of aspects of systems. They may have partial knowledge of some of the definitions of system terms but cannot use them in a consistent nor strongly coherent manner. While they can identify outcome variables (e.g. stocks that are explicitly part of the goal state), they are not able to track a causal link and they largely focused on macro-level directly observable variables. Their predictions and explanations are a-causal; i.e. more assertions than any cause and effect relations (e.g. "things happen because that's the way they are" Brown, 2005, p. 7).
2	Students have an elemental understanding (Brown, 2005, p. 7) of some aspects of systems – they can use models to represent simple, single cause and effect relations but without strong justification i.e. they are still prone to common misconceptions, e.g. they tend to only relate macro-level, directly observable causes and effects rather than identifying hidden variables and factors. This is due in part to not being able to understand and analyze a system at different levels (Cheng et al., 2010.) They are better at explaining than predicting.
3	Students have a locally coherent understanding of many aspects of systems. Students can use system thinking terms to describe components and system relations in some contexts and use different representations. They can use models to represent bivariate cause and effect relations along with strong justifications. They can relate binary combinations of hidden and directly observable combinations, and even single causes to multiple effects. i.e. they are less prone to common misconceptions but still are limited linear thinking with single causes (which may or may not be chained together.) They have a rudimentary understanding of negative feedback and can use it to explain and predict change in behavior of a system over time. They still are not able to consistently understand and analyze a system at different levels (Cheng et al., 2010.)
4	Students can relate multiple causes to multiple effects as long as they behave in simple ruleful ways (e.g. cases in which all causes are needed for the effect to occur, cases in which all causes contribute independently to the amount of the effect as in Pollution City, etc .i.e. the causes are not emergent but are instead explainable in terms of the causal component parts. This level is consistent with Brown's (2005) conceptual depth level 4. Students can apply this scope of understanding within a wider range of contexts than in prior levels.
5	Students have a globally coherent understanding of many aspects of systems thinking in many contexts. They can analyze of moderately complex system that includes multiple variables, including several hidden variables, feedback spread out in space and time, and emergent behaviors that requires understanding a system at multiple levels, with multiple causes interacting to create complex emergent effects (corresponding to level 5 in Brown, 2005).

In Jackson City, in order to be able to assess aspects of students' systems thinking we need to put them in situations in which a complex system is governing what is visible, they can take actions which affect its behavior, and they are motivated to actions that change the system's behavior in targeted ways. All of this is to provoke systems thinking in students and consequent actions on their part that give us clues about that thinking. This sounds like data of the first kind, student actions. This is so in the sense that student actions are in the foreground. We will see, however, that the "observable variables" in assessment always concern contextualized actions, or meaningful observations of action in light of features of situations and sometimes additional information.

Figure 6:
An Assessment Design Argument

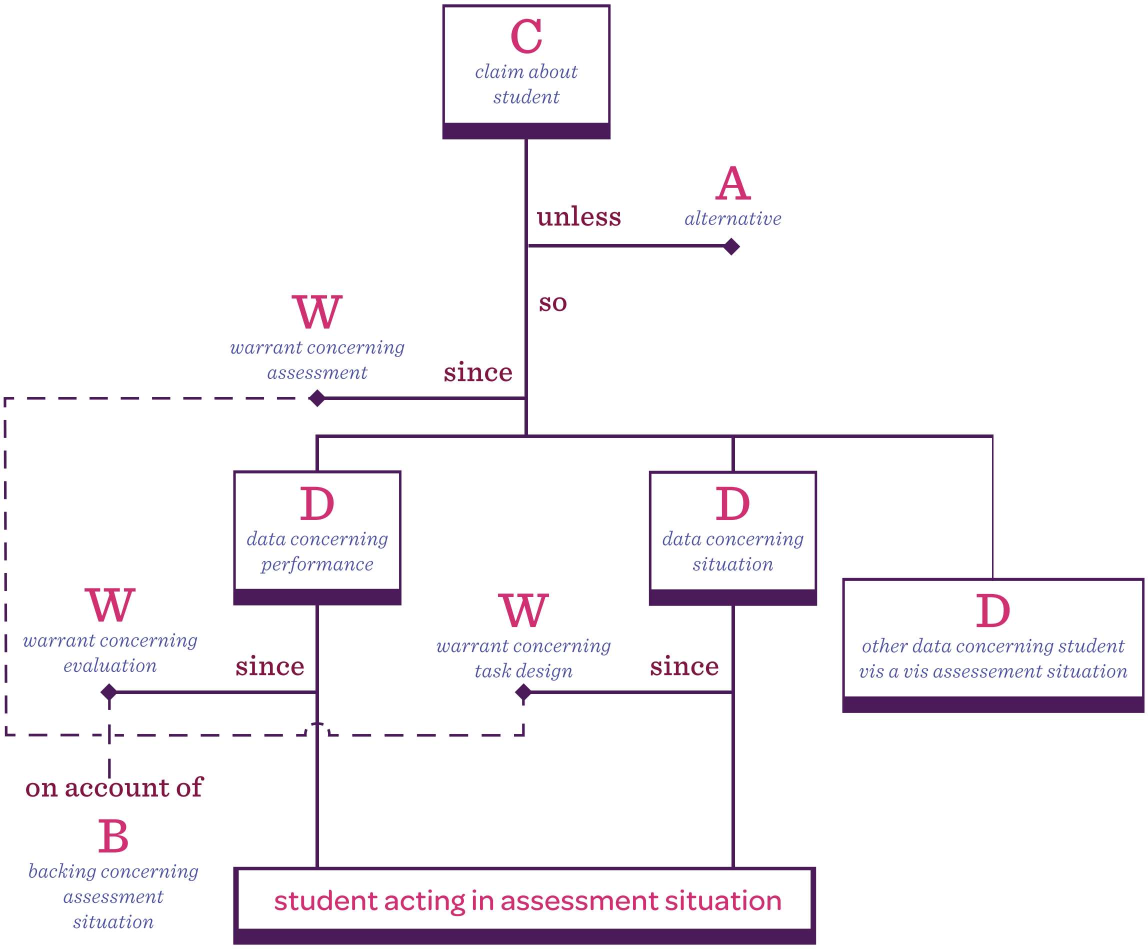

C
claim about student
unless
A
alternative
so
W
warrant concerning assessment
since
D
data concerning performance
D
data concerning situation
D
other data concerning student vis a vis assessement situation
W
warrant concerning evaluation
since
W
warrant concerning task design
since
on account of
B
backing concerning assessment situation
student acting in assessment situation

Figure 6
Adapted from Figure 2 from Mislevy, R.J. (2005). Issues of structure and issues of scale in assessment from a situative/sociocultural perspective. CSE Technical Report 668. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Copyright 2003 by The Regents of the University of California. All rights reserved.

The third kind of data that an argumentation perspective highlights is "what else is known." As a simple example, what we infer from a student's correct use of fraction subtraction is quite different if we know it is just like the problems she worked on in class yesterday, if it is a novel application of ideas from a recent lesson, or if she has never worked with fractions before but figured it out on her own. This is important in GBA for a number of reasons:

- It affects whether potential alternative explanations can be ruled out, for example by knowing what prior knowledge a student coming to a GBA, or what supports and hints might be required, or whether the student is familiar with some aspects of the content and the reasoning challenges in a GBA but not others. Features that may not seem at first to have to do with assessment can have substantial impact on the evidentiary value of observations.

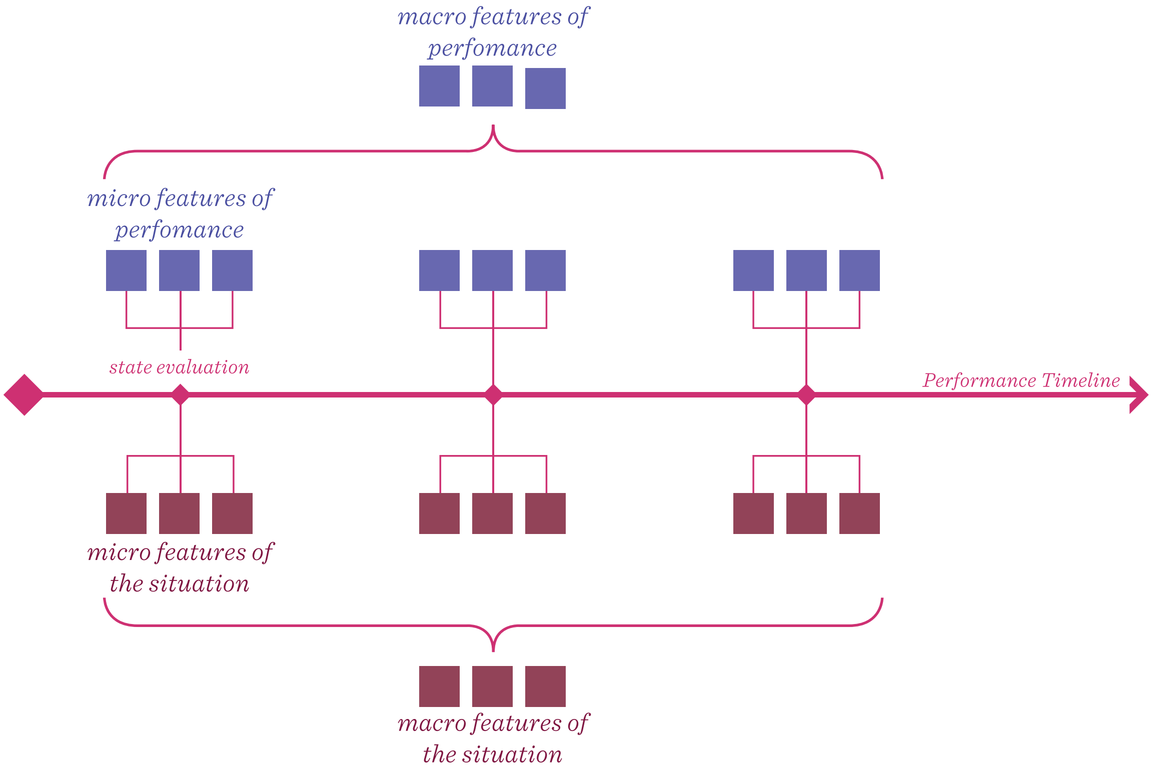
- Knowledge and skills that are not germane to the assessment's purpose yet are required for successful performance need to be either supported so it is known, or included as nuisance variation in psychometric models. This is why using a GBA as part of a course simplifies the assessment modeling challenges and provides more useful information. It is not the game alone that determines the evidentiary value of the observations, but also the standpoint of knowledge of the user. The user can be one or more of these possibilities: the GBA itself, the student, the teacher, or a distant party such as a chief state school officer. Generally, the more distant users are from the context, the less of this additional information they have, the more alternative explanations they must entertain, and hence the less evidence the data provides to them.

- Additional information can also condition how features of a student' actions are interpreted and how features of the situation are interpreted. An example that is simple yet critical for inferences about learning transfer is whether a task is, for a given student, the same one she has already worked with, similar to a familiar one, or novel. The data of this type that we use in the evidentiary argument is neither in the features of the task per se nor of the student's learning history but in their relationship to each other.

- Whereas items in traditional tests are independent, performance across situations in games and simulations generally has serial dependence. That is, what a student does at one point affects the situation that arises, and can impact the interpretation of both the situation and the student's action in it. For example, checking whether a value is providing the right output might usually be a good move in fixing a hydraulic system, but it is not a good move if an examinee has already seen that there is no flow in the hose leading into it. Features of the current and sometimes previous situations and actions might be necessary to make sense of either. Figure 7 shows the ways that identification of features of situations and actions may need to be determined in continuous-activity assessments.

In games, this kind of information can be captured and stored in a finite state machine—a set of variables whose values at a given point in time define the state of the game, and determine what happens next, what options are available to players, capabilities of characters, goals achieved and yet outstanding, and perhaps hundreds or other aspects of the situation. In Chapter 5 we saw that game designers often use finite state machines to organize complex game play. We can do the same to manage data for assessment in a GBA.

Figure 7:
Detail of Features of Performance and Situation in an Interactive Task



macro features of perfomance
micro features of perfomance
state evaluation
Performance Timeline
micro features of the situation
macro features of the situation

Figure 7
Adapted from Figure 3 from Mislevy, R.J. (2011). Evidence-Centered Design for Simulation- Based Assessment. CSE Technical Report 800. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Copyright 2011 by The Regents of the University of California. All rights reserved."

The argument structures we just described are useful in thinking about what to build into play situations in GBA so they will provide evidence, but they don't offer specific guidance for particular learning goals. Another Domain Modeling representation called an Assessment Design Pattern organizes substantive information from Domain Analysis into categories that are related to the elements of assessment arguments. Mislevy, Riconscente, and Rutstein (2009), for example, provide a suite of design patterns for building assessment tasks for model-based reasoning in science--building models, reasoning through them, critiquing them, revising them, carrying out investigations with them. Standards documents such as the NGSS can be starting points for design patterns.

A design pattern is centered around some aspect of reasoning or some "big idea" in a subject provides support for task developers in several ways. It describes features that situations will need to have in some way in order to get evidence about the focal capabilities, other features that can be varied to adjust difficulty, other knowledge and skill that might be involved, kinds of things students can say or do to provide evidence and what aspects of them are important to look for. Design patterns are written at a level of generality to make them useful for performance tasks, simulations, and GBAs as well as familiar assessment tasks. The designers of Jackson City drew in part on the design pattern for systems thinking in Cheng et al. (2010).

Adapted from Figure 2 from Mislevy, R.J. (2005). Issues of structure and issues of scale in assessment from a situative/sociocultural perspective. CSE Technical Report 668. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Copyright 2003 by The Regents of the University of California. All rights reserved. The variant of design patterns we are using in specifically in GlassLab is illustrated in Table 3, which gives guidance for developers on kinds of situations and actions which, in the course of game play, will provide evidence to ground claims about a player's current level of systems thinking relative to her actions in the current game context.

The design pattern starts to give a meaning to Systems Thinking as it will become operationally defined in Jackson City. The left column starts to add clarity to the meanings of the levels defined briefly and abstractly back in Table 2, still broadly enough to apply to contexts other than Jackson City. They could be used as a coherent framework for designing challenges and conducting assessment in GBAs that foster systems thinking with different systems and contexts. The right column in Table 3 is specific to the Jackson City design space. It motivates more particular design features for situation features, player affordances, and evaluation approaches that will be addressed in the Conceptual Assessment Framework layer discussed next.

Whether reasoning at, say, Level 3 in Jackson City has anything to do with the kinds of reasoning a player might do in another context or a different system is quite another matter, as we will discuss in the Validity section in Chapter 12. The answer can depend on the way that players are led to develop their understanding across contexts, and facilitate their learning in new contexts (Bransford & Schwartz, 1999). This issue highlights the distinction between the variables we will be using as pieces of machinery to manage evidence in particular contexts, and psychological meanings they might merit

Table 3:
Design Pattern for Systems Thinking

Distinctions Between Levels	Potential Evidence
Level 1: Fragmented understanding **At this level students can:** Define or identify some specific definitions of systems terms in text. **At this level it will be hard for students to:** Construct models that relate two variables a direct causal link with no feedback (one stock, one flow) e.g. "pollution is caused by the smoke coming out of the coal power plant." Create coherent system models that include multiple causal links or chains.	In integrating information, selected text does not align with game play nor fit in diagram **Puzzle solutions:** 1. Unsolved 2. Solution path: actions do not match known causal links 3. Little or no exploration of data **In causal loop diagrams:** 1. Only a small subset of relevant variables selected 2. Focus on assertions rather than causal links or links do not match reality or anything systematic
Level 2: Elemental understanding **At this level students can:** Solve puzzles involving one or two observable macro-level variables. **At this level it will be hard for students to:** Understand, model, and predict behaviors involving small numbers of variables with some hidden variables, or negative feedback to reach a stable equilibrium. Cannot solve puzzles that involve reasoning to relate multiple levels of system behavior (micro, macro-emergent)	In integrating information, selected text fills out diagram (quantities, dynamics) but not justification **Puzzle solutions:** 1. One cause to one effect only 2. Solution path: search for macro level observables 3. Solution path: exploring data for directly observable variable **In causal loop diagrams:** 1. Most selected variables are relevant but can be missing some 2. Only single cause to single effect links 3. If text tool used, little evidence selected and/or justification not well formed
Level 3: Locally coherent understanding **At this level students can:** Solve puzzles that involve one or two observable macro-level or hidden variables, and simple linear chains of variables, and rudimentary negative feedback loops. **At this level it will be hard for students to:** Understand, model, and predict behaviors involving multiple causes that create single or multiple effects. They will have difficulty solving puzzles that involve reasoning to relate multiple levels of system behavior (micro, macro-emergent) and reasoning chains that go beyond linear chains.	In integrating information selected text fills out diagram (quantities, dynamics) and includes reasonable justification. **Puzzle solutions:** 1. Solution (solved) with substantial scaffolding 2. Solution path: search for macro and micro level observables and hidden variables 3. Solution path: exploring data for directly observable variables and hidden variables **In causal loop diagrams:** 1. Selected variables are a good match to the full set of relevant variables 2. Links are single cause to single effect 3. Selected justifications are strong within scope of single cause/effect links selected
Level 4: Multiple causes to multiple effects **At this level students can:** Solve medium difficulty SimCity puzzles involving multiple variables and feedback as long as there is no emergent behavior. Students can express their solutions using system terms, and with system diagrams, abstracting from a variety of specific contexts to general models. **At this level it will be hard for students to:** Solve especially complex puzzles with tricky feedback relations, and complex emergent behaviors.	**Puzzle solutions:** 1. Solution (solved) with little scaffolding as long as effects are explainable by analysis of individual components rather than being emergent 2. Solution path: search for macro and micro-level observables and hidden variables 3. Solution path: exploring data for directly observable variables and hidden variables **In causal loop diagrams:** 1. Selected variables match the full set of relevant variables with no extras 2. Links include multiple cause to single effect as warranted 3. Selected justifications are strong & explain full set of cause/effect links
Level 5: Globally coherent understanding **At this level students can:** Solve moderately complex SimCity puzzles involving multiple variables, feedback, and the need to explain emergent behavior. Express their solutions using system terms, and with system diagrams. **At this level it will be hard for students to:** Solve especially complex puzzles with tricky feedback relations, and complex emergent behaviors.	**Puzzle solutions:** 1. Solved with little scaffolding even if the effects are emergent 2. Solution path: search for macro and micro-level observables and hidden variables 3. Solution path: exploring data for directly observable variables and hidden variables **In causal loop diagrams for systems including emergent effects:** 1. Selected variables match the full set of relevant variables with no extras 2. Links include multiple cause to single effect as warranted 3. Selected justifications are strong and explain full set of cause/effect links

Student, Evidence, and Task Models

The layer of ECD called the Conceptual Assessment Framework or CAF contains specifications for the more technical objects that constitute an assessment. The three main ones are student-, evidence, and task models. It is in the student and evidence models that the machinery of psychometrics is specified, but it is grounded on the assessment argument structures wrestled out (at least provisionally) in conversations at the Domain Modeling layer. Those conversations will have looked forward to psychometric modeling issues, if not specific forms, in the same way that they looked ahead to game mechanics, encompassing the viewpoints of both design domains.

Figure 8 gives a high-level representation of the main CAF models. Internal structures can be detailed in various ways, as described for example in Mislevy, Steinberg, and Almond (2003) and Riconscente et al. (2005) (also see Luecht, 2003, and Gierl & Lai, 2012). The following paragraphs discuss the kinds of things they contain, note how they are used in familiar assessments, and look ahead to their roles in GBA. The section on delivery layer will say more about how these specifications take form in operational elements of an assessment and how they shape activities, internal messages, and external reports.

Figure 8:
The Central Models of the Conceptual Assessment Framework

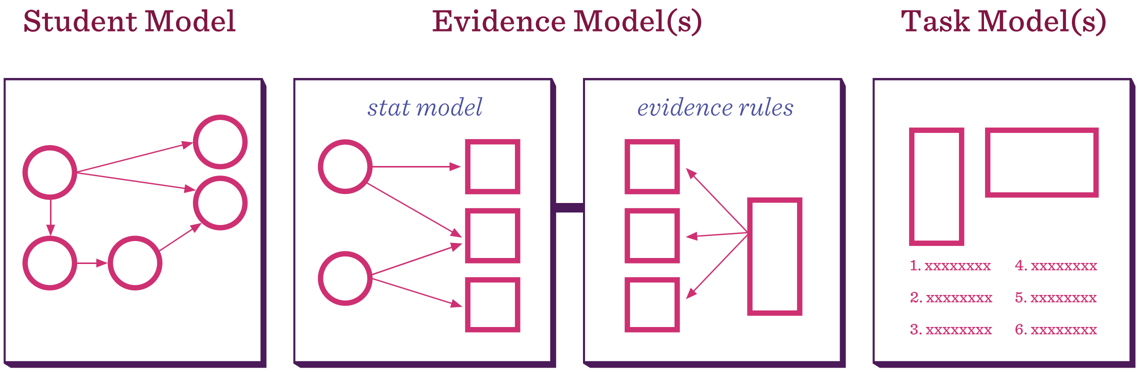


Figure 8
Based on Figure 1 from Mislevy, Robert J.; Almond, Russell G.; Lukas, Janice F.(2003)
ETS Research Report RR-03-16 "A Brief Introduction to Evidence-Centered Design." Reprinted with permission.

The Student Model at the left of the figure contains variables for expressing claims about targeted aspects of students' knowledge and skills. These are student model variables, or SMVs. SMVs are formalizations of aspects of students' capabilities that are needed to express the claims in assessment arguments.

The number and character of SMVs in an assessment depends on the purpose(s) of the assessment, or, in interactive assessments like GBA and intelligent tutoring systems (ITSs), the purpose(s) at some given level in a hierarchy of purposes. In simple familiar assessments, the student model consists of one variable, an overall proficiency in a domain of tasks, which is reflected in some kind of summary score. GBAs and ITSs generally need to track multiple aspects of proficiency, which may be involved in different ways in different situations, and which may change as students interact with the system (e.g., they learn). Proficiencies cannot be observed directly; we must use what we see students say or do to base inferences about proficiencies on. Key issues in any application of psychometrics are what the nature of these student model variables ought to be. What they actually turn out to be -- that is, their situated meanings -- is determined by the patterns in the students' actions in the situations we design for them to act in.

Table 2 was a substantive precursor for a student-model variable for systems thinking in Jackson City. This is a kind of entity one works with to start to build assessment arguments. To implement an argument in the elements and processes of an operational assessment requires embodying the concepts in "pieces of machinery." A student-model variable in this sense is a variable that can take on different values, such that at a given point in time and a given standpoint in information, a probability distribution can indicate degree of belief about possible values. The construct suggested in Table 2 could be implemented as a discrete variable with five levels, or as a continuous variable of increasing capabilities, with regions corresponding to the levels suggested in the table. The former choice will be illustrated here, with a five-level ordered variable called SystemModeling.

At the right of Figure 8 is the Task Model. It describes salient features of assessment (game) situations, in terms of task-model variables which can take different values. They encompass characteristic and variable features of situations that were learned in Domain Modeling. In familiar assessment, these are forms and descriptors of distinguishable, well-defined, tasks. In simulations, performances, and GBAs, "tasks" need not be predefined, but can also be recognized as evidence-evoking situations that arise as the student acts in the environment.

Task models also include specifications of Work Products, or what is captured of what students say or do. In familiar assessments, these are discrete response or captures of performances such as essays or problem solutions. These can be required of GBA players as well, but GBAs can also capture more detailed records of evolving game states, students' interactions (variously called log files, click streams, slime trails, and transaction lists), and even eye-movement traces, facial expressions, and physical measures such as respiration and squirming. Chapter 8 will look more closely at work products in GBAs.

Evidence models are the bridge between what we see students do in various situations (as described in task models) and what we want to infer about students' capabilities (as expressed in student-model variables).

- The *evaluation component* says how one identifies and evaluates the salient aspects of work products, expressed as values of Observable Variables. For multiple-choice items, the evaluation component just compares a student's answer with the correct answer, and returns 1 or 0 as the value of the observable variable. More complex features may need to be interpreted from more complex work products or relationships among them (e.g., efficiency of a solution, whether a patient was stabilized, how systematic an investigation is). Carrying out such evaluations can require the use of meta-data (what else is known a priori about the student and the tasks, such as student background and task features) or paradata (contextual data that accompany response data, such as features of the situations and the consequences of the student's previous interactions with the system; instructional designers and data miners have adopted the term from the sample survey literature.). This component embodies the reasoning in the assessment argument from performance to features of performance. This component contains information needed to implement the *Evidence Identification* process in the delivery system layer. It is the subject of Chapter 9.

Table 4 and Table 5 summarize two Observable Variables from Jackson City we will use as examples, MultivariateThinking and JobsPollutionEndstate. The discussion in Chapter 9 of identifying variables from log files will say more about the procedures by which their values are determined—in this case rule-based evaluation of features of players' actions. MultivariateThinking is determined from a fixed task players must complete, creating a diagram of the interrelationships among factors affecting pollution and jobs in Jackson City. JobsPollutionEndstate is determined from the less constrained series of actions they take as they modify the city to reduce pollution while retaining jobs.

- The *measurement model component* contains statistical psychometric models that synthesize data across situations, in terms of updated belief about student-model variables. The simplest measurement models are classical test theory models, in which scores based on observable variables are added. Modular construction of measurement models assembles pieces of more complicated models such as those of item response theory or Bayesian inference networks (e.g., Mislevy et al., 2002). This component contains information needed to implement the *Evidence Accumulation* process in the delivery system layer. Chapter 10 looks more closely at measurement models.

Table 4:
Levels of MultivariateThinking (Observable / Variable)

Value	Description
0	Not comprehensible.
1	Use all given factors, add none of their own. No indications of factors relating to each other or having anything other than direct relationship to jobs/pollution.
2	Added some user-generated factors. Still no indications of factors relating to each other or having anything other than direct relationship to jobs/pollution.
3	At least some factors are related in some way to both jobs and pollution. No indirect relation relationships or negative relationships.
4	Indication of indirect and/or negative relationships.

Table 5:
Indication of Indirect and/or Negative Relationships (Observable / Variable)

Value	Description
1	Pollution remains high, jobs very low.
2	Pollution reduced somewhat from initial levels high, jobs very low.
3	Performance reflects mostly univariate thinking on either jobs or pollution, given observed gains on one dimension, but lacking in the other.
4	Performance reflects mostly univariate thinking on either jobs or pollution, given observed gains on one dimension, but lacking in the other.
5	Clear improvement on pollution and strong evidence for multivariate reasoning to maintain or improve the jobs condition.

Implementation

The Assessment Implementation layer of ECD is about constructing and preparing the operational elements specified in the CAF. In familiar assessments, this includes authoring tasks, finalizing rubrics or automated scoring rules, and estimating the parameters in measurement models. Using common and compatible data structures increases opportunities for reusability and interoperability, and helps bring down costs. For discussions in the context of simulation-based assessment, see Chung, Baker, Delacruz, Bewley, Elmore, and Seely (2008) on task design; Mislevy, Steinberg, Breyer, Almond, and Johnson (2002) on measurement models; Luecht (2009) on authoring frameworks; and Stevens and Casillas (2006) on automated scoring.

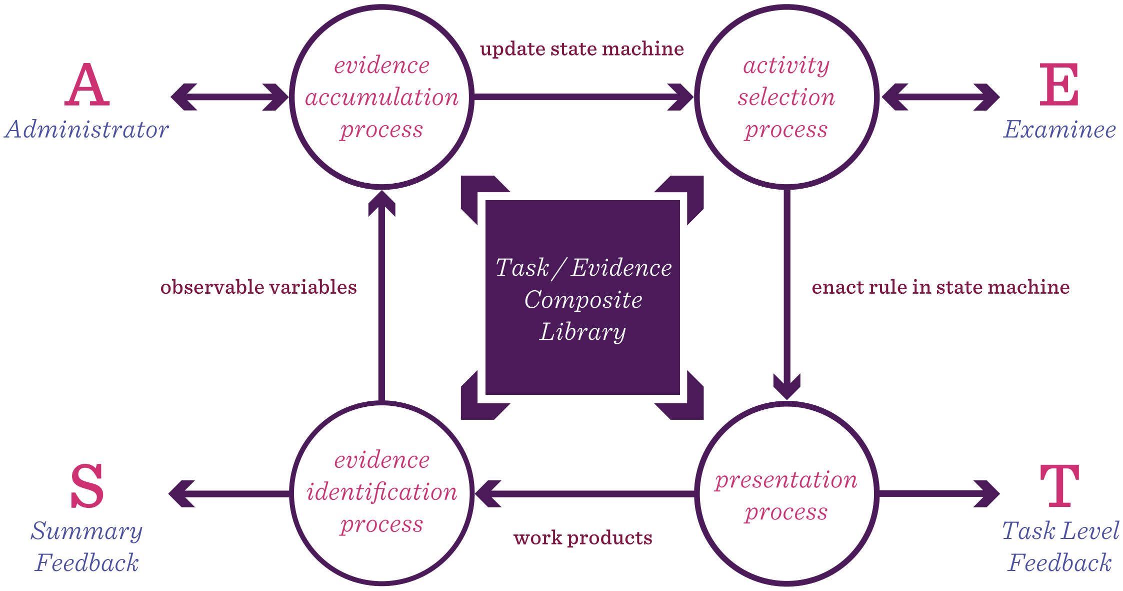
In GBA, efficiencies can be gained with respect to both gaming and assessment by exploiting re-usable elements, both conceptual and "mechanical." With regard to assessment in particular, we have already mentioned design patterns as a way of organizing thinking about ways to elicit and capture evidence about recurring aspects of students' proficiencies, and they are most useful in areas that are at once hard to assess and well-matched to games, such as systems thinking, investigation, using representations, and building and using models. Re-usable pieces of machinery that can be applied across content areas include adaptable structures of presenting and capturing information in the game environment (drawing for example on Scalise and Gifford, 2006); processes and structures for identifying evidence in log files and defined work products; and modular structures and general processes for psychometrics (e.g., Mislevy et al., 2002).

Game design has its own armamentarium of re-usable elements. Chapter 13 will discuss the idea of creating GBA building blocks that combine aspects of game experience from the play perspective and evidence elicitation from the assessment perspective.

Assessment Delivery: The Four-Process Architecture

The Assessment Delivery layer concerns the processes, messages, and calculations that occur when students actually interact with assessment situations, their performances are evaluated, and feedback and reports are produced. Almond, Steinberg, and Mislevy (2002) lay out a four-process delivery system that can be used to describe not only computer-based testing procedures, but paper-and-pencil tests, informal classroom tests, tutoring systems, and game-based assessments. When an assessment is operating, the processes pass messages among themselves in a pattern determined by the test's purpose. All of the messages are either data objects specified in the CAF (e.g., parameters, stimulus materials) or produced by the student or other processes in data structures that are specified in the CAF (e.g., work products, values of observable variables). Common language, common data structures, and a common partitioning of activities again promote the reuse of objects and processes, and interoperability across projects and programs. Figure 9 shows the four principal processes.

Figure 9: Processes in Assessment Cycles



A
Administrator
evidence accumulation process
update state machine
activity selection process
E
Examinee
observable variables
Task / Evidence Composite Library
enact rule in state machine
S
Summary Feedback
evidence identification process
work products
presentation process
T
Task Level Feedback

Figure 9
Based on Figure 2 from Mislevy, Robert J.; Almond, Russell G.; Lukas, Janice F.(2003)
ETS Research Report RR-03-16 "A Brief Introduction to Evidence-Centered Design." Reprinted with permission.

- The *activity selection process* concerns what to do next. In a test, it selects a task or activity from the task library, or creates one in accordance with templates in light of what is known about the student or the situation. In a GBA, it is subsumed in one or more finite state machines at some level that govern game activity, and may use current knowledge about the student as one of the variables being tracked and utilized.

- The *presentation process* controls interaction with the student. It is responsible for presenting the task to the student, managing the interaction, and capturing work products. As mentioned, there can be hierarchical nesting for managing the interaction. That is, there may be a four-process cycle like Figure 9 for a multi-level game, but within the presentation process at a given level, a finer-grained four-process system for that level, and perhaps even finer-grained ones within them.

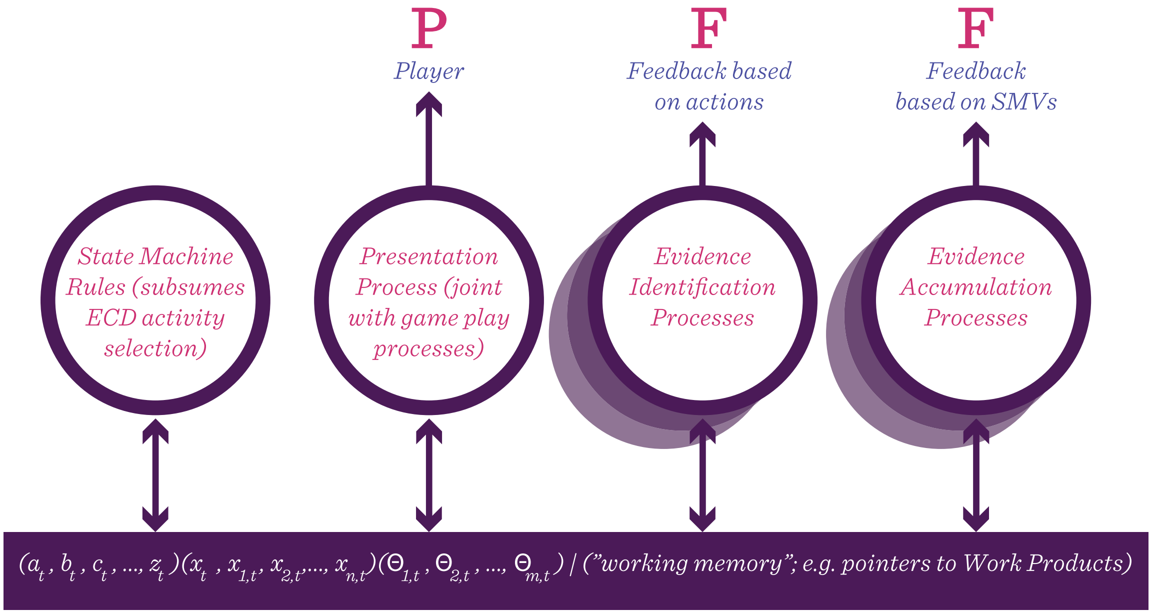
- Work Products are passed to the *evidence identification process* (called task-level scoring in familiar assessments). It evaluates work using the methods specified in the Evidence Model. It can send the resulting values of Observable Variables to the *evidence accumulation process*, and to the *activity selection process* to provide more immediate feedback based on what the student has done, such as hints or comments. We will call these two uses of observables inferential feedback and task level feedback.

- The *evidence accumulation process* (called test-level scoring in familiar assessments) uses measurement models to summarize evidence about the student model variables (for inferential feedback) and produce score reports or, if the assessment is simulation-based or a GBA, to update the finite state machine that the Activity Selection Process represents.

A fixed-form multiple-choice test requires a single trip around the cycle. A simulation-based task or a GBA can require many interactions among the processes in the course of a performance. And, as mentioned, there can be hierarchies of this structure: For situations within scenarios, for scenarios within levels of a game, and for game levels. For example, an intelligent tutoring system can jump out to instructional or practice modules (Shute & Psotka, 1996), which can be viewed as particular kinds of "tasks." All of these kinds of interactions among assessment processes are based on the same game and player interactions that constitute game play. As mentioned, most games manage these interactions with finite state machines. Figure 10 shows the four kinds of assessment processes in relation to a finite state machine in a GBA. Note that a GBA can contain multiple evidence identification and evidence accumulation processes.

We will say more about how aspects of game-play activity in GBAs undergo processing that play roles in assessment. One point of some interest will be "telemetry" (game play information gathered from the players' computer and sent to the central server) in games where processes reside in different locations (on the user's local device, linked devices across users, on a central server).

Figure 10:
The Four Assessment Delivery Processes and a GBA Finite State Machine


P
Player
F
Feedback based on actions
F
Feedback based on SMVs
State Machine Rules (subsumes ECD activity selection)
Presentation Process (joint with game play processes)
Evidence Identification Processes
Evidence Accumulation Processes
$(a_t, b_t, c_t, ..., z_t)(x_t, x_{1,t}, x_{2,t},..., x_{n,t})(\Theta_{1,t}, \Theta_{2,t}, ..., \Theta_{m,t})$ | ("working memory"; e.g. pointers to Work Products)

The Interplay between Design and Discovery

We have now reviewed fundamental ideas in both game design and assessment design. Chapter 13 will say more about actual design processes for GBAs. A few comments are in order at this point in the discussion, though, before we discuss a number of more technical psychometric components.

Design is almost always iterative. We design artifacts prospectively as well as we can, taking account of principles, exemplars, and practical experience. However initial thoughts are rarely final thoughts. Things don't fit together exactly as expected, clients shift their priorities, users don't perceive features as they thought they would, and materials, timelines, and budgets force unanticipated revisions.

Iterations are expected in GBA design, if for no other reason than game design and assessment design are already both iterative on their own. Game designers in particular use processes with frequent user tests and revisions. These are called "agile" design processes, in contrast with so-called "waterfall" design processes that are designed and implemented, stage by stage sequentially, with relatively fewer and more widely spaced testing. Agile design processes are especially well suited to new products, such each new games. Very large changes are not at all unusual early in the process, so that the version of a game that is released may bear little resemblance to early rapid, inexpensive, prototypes.

Familiar kinds of assessments have iterations and testing as well, but fewer and less frequent, and are more like waterfall design processes. This is especially so when a planned assessment is planned that is very much like very many previous assessments. It is almost possible to march through the levels of the ECD framework as though they were stages of production, with a few feedback cycles for reviews and pretesting. Items, instructions, timing, presentation details, and scoring rubrics are fine tuned, but rarely are there major changes in fundamental goals of measurement, test specifications, forms of evidence, and psychometric models.

More revisions are typical, however, in more complex assessments such as GBAs, interactive simulation tasks, and hands-on performances. Like games, more is new and more is going on; more elements can interact with one another in unexpected ways; and more can go wrong in both design and evaluation. Designing a complex assessment is more like designing a game: Based on what one knows early on, expressed loosely in terms of sketches of assessment arguments or rough design patterns from initial domain analysis, successive prototypes are tested earlier and more often. More fundamental changes can occur at different levels and in different places of the ECD framework:

Because performances can be complex, scoring methods are open to greater exploration and discovery. One bootstrapping method is to include some work products and observable variables that are fairly well understood, and use these as anchors in supervised learning to seek patterns in performance that can provide additional evidence.

Unlike familiar assessments, it is possible to revise or augment the set of student-model variables. Exploratory data analysis (particularly visualization and hypothesis generation tools) and educational data mining techniques (including recent methods such as unsupervised neural network modeling and natural language processing tools, as well as long-established psychometric methods such as factor analysis, cluster analysis, latent class analysis, and multidimensional scaling) can identify associations among observable features of performance that suggest new student-model variables.

Insights from data mining with respect to both observable variables and student-model variables can suggest in turn improvements to the design of situations and affordances. These can produce stronger evidence by identifying sources of construct irrelevant demands to reduce by redesign or support; providing additional opportunities in the situation to observe the newly discovered forms of evidence; and, if appropriate, to evoke more directed and structured forms of the new evidence.

Such opportunities present themselves more in GBAs than assessments. The lower stakes associated with games meant to support learning can generate large amounts of data for exploration, and because accurate comparisons are not required for students at different times points, new releases of a GBA can continually improve assessment as well as game play. Among big data domains, games have the advantage of being able to feed insights back into improved design, with the improvements for assessment always structured around bolstering the assessment argument (Bennett & Bejar, 1998). Design and discovery intertwine through rapid iterative cycles.

Orientation to Psychometrics in GBA

In common usage, the term "psychometrics" includes activities that ECD distinguishes as evidence identification and evidence accumulation. This short chapter describes the nature and roles of the two processes in assessment reasoning, with an eye toward their use in game-based assessments. The following chapters will look more closely at work products, evidence identification processes, and evidence accumulation processes.

Evidence Identification concerns reasoning from particular observed performances to values of observable variables—that is, identifying features of the performance or product that are data for an assessment argument, as nuggets of evidence about students' capabilities.[3] The reasoning runs in one direction: From particular realized work products, to particular values of observables to characterize the salient features of the performance. Of the two processes, Evidence Identification is more intuitive, even when the methods by which it is produced are quite esoteric (as in natural language processing of essays).

Evidence Accumulation is modeling probability distributions of these observable variables as functions of aspects of students' knowledge, skill, propensities, or other more extensive characteristics of students. These are expressed as unobservable or latent variables in psychometric models. It is they, rather than the specific performances themselves, that are targets of learning and targets of assessment. In ECD terminology, these are the student model variables (SMVs) or proficiency variables.

This way of modeling is not familiar to most people. It is worth pausing to say just how it differs from the more comprehensible (even if complicated) evidence identification processes, how the two work together, and their advantages in assessment generally and at least some points in at least some GBAs.

The key idea is that student model variables express tendencies or capabilities that can be used to model performance across multiple situations, actual or hypothetical. We don't observe them directly, but we can make inferences about them based on the particular things students do. Their values might be assumed to remain constraint over some period of observation, as in familiar large-scale tests, or they might be presumed to change over time through experience, as in instructional systems. Either assumption might be appropriate for a given GBA, and there can be mixes across different SMVs. For example, an SMV for reading proficiency used to tune feedback to a player could be assumed constant over the course of a game, but SMVs for systems thinking, which the game is designed to improve, are modeled as changing as the game progresses. There can also be a blend for

[3] Data is just "stuff." It doesn't become evidence until we establish its relevance in some inference (Schum, 1994). The same data can be strong evidence for one inference, weak evidence for another, and none at all for a third.

given SMVs. For example, some SMV values can be approximated as constant within small segments of a game, with changes modeled when a player moves to the next segment (see Kimball, 1982, for an early application of this idea in an intelligent tutoring system).

There are several reasons to consider using psychometric measurement models and SMVs in assessments in general, and in GBAs in particular, for at least some kinds of inferences. (Immediate feedback based on aspects of specific performance can be useful for hints and observations, and both can be done in the same system.) Among ones we can take advantage of are the following:

- Psychometric measurement models transform data about specific performances into evidence for beliefs about characteristics / propensities / proficiencies of students, in terms of SMVs (including ones that are changing as the game progresses). They can synthesize evidence from disparate forms of data across specific situations, in terms of more underlying aspects of students' knowledge, skills, identities, values, and epistemologies (SKIVE elements, as Shafer, 2006, calls them).

- The SMVs are of more persistent interest than particular actions, and are directly connected to educational frames such as standards and learning progressions.

- The way psychometric measurement models are built—conditional probability distributions for observable variables given potential configurations of SMVs—puts them in the world of probability models, which grants us access to five centuries of insights, research, and methodology. Specifically, we can adapt conceptual and statistical tools from psychometrics. The following properties all flow from this one.

- We know how to build models that account for complexities such as multiple aspects of proficiency being involved in various mixes for different aspects of performance, dependences introduced by time and problem structures, different forms of data, and changing values of SMVs as students learn. Not to say that this is easy in any given application, but there are well-understood logics and models for doing so.

- Once the models are in place, we know how to update beliefs about a student's SMVs as evidence arrives (specifically, through Bayes theorem). We can do this sequentially or in batches, and take into account the complexities noted above.

- We know how to characterize weight and direction of evidence. It will be in terms of indices for properties such as reliability, standard error of measurement, and classification accuracy. This gives designers metrics to evaluate the effects of design decisions about the game with respect to evidence, as well for engagement or learning, and to characterize and compare tactics for evidence identification.

- The probability models for different kinds of sources of evidence can be modular, assembled in real-time, and adapted to design changes for parts of a game without needing to revamp a given psychometric framework.

The signal achievement of psychometrics as methodology since its origins at the beginning of the 20th Century was made by "treating the study of the relationship between responses to a set of test items and a hypothesized trait (or traits) of an individual as a problem of statistical inference" (Lewis, 1986). This move allows the formidable tools of probability based reasoning to be brought to bear in assessment reasoning. The challenge today is to extend these accomplishments from applications to relatively sparse and encapsulated data, for inferences cast in trait and behaviorist psychology, to the richer data made possible in interconnected digital environments, for inferences cast in contemporary sociocognitive psychologies, as encountered for example in game-based assessment (DiCerbo & Behrens, 2012).

Capturing Data: Work Products in GBA

The epiphenomena[4] of a student playing a game-based assessment are multiple and diverse, and unique to each instance. Most obvious are mouse clicks and keystrokes at certain times in certain places, as something is happening on the screen and from the speakers. Behaviorally, the player moves, talks, sweats, grimaces, and laughs. Moving further inward, her heart-rate, breathing, skin conductivity, capillary dilation, and eye fixations all fluctuate as she plays. Electrical activity dances in her brain, hormones ebb and flow through her bloodstream. These are all aspects of the "performance in a situation" box at the bottom, the starting point, of the assessment argument (Figure 6). Our interest in learning and assessment is the cognitive structures and activity patterns which, in continuous interaction with the game, players assemble in order to act, and which give meaning to all of these goings-on.

This chapter looks more closely at capturing data in GBAs as a source of evidence in assessment arguments. From the game perspective, we are looking at the many things that players do in various situations in the game, as they act to achieve their goals in the various situations they encounter. In ECD terminology, we now think of captured forms of some of the performance in terms of the work products—such as structured things students create or properties of those things, problems they solve or steps by which they solve them, places they visit, how they get there, and in what orders they take which actions at what times. Clearly games can generate huge volumes of data, in the form of clicks, actions, time stamps, and concomitant game conditions. For assessment, the trick is figuring out what among all this data is evidence. We will focus on data within the small-g game.

Before discussing the kinds of work products we might capture in GBAs, we can note some levels for successive perspectives on data about players' performances. Defining work products is the key first step, but knowing where we are going in evidence identification and evidence accumulation processes sheds light on design decisions that need to be made about work products.

Table 6 provides some imperfect but useful language for talking about reasoning from evidence in GBAs. A GBA is designed to produce epiphenomena (Level 1) that will eventually give us information about Proficiency structures (Level 5). We use knowledge about their grammar (Level 2) and about the game, content, and purpose to define Work Products. Work Products capture distillations of the vast and sundry epiphenomena in some form, perhaps still massive, but already selected, filtered, and focused to some degree to be a higher grade ore for nuggets of evidence. Those nuggets take the form of meaningful features in work products (Levels 3 and 4); this step is Evidence Identification process in ECD. (Figure 11 shows meaningful interpretations of raw actions in Jackson City.) These features

[4] "A secondary phenomenon accompanying another and caused by it." Merriam-Webster - The Free Dictionary. Accessed 7/31/2013 at http://www.merriam-webster.com/dictionary/epiphenomenon "There must be a pony in there somewhere."

are still characterizations of performance, but more heavily interpreted. The student model variables in psychometric models (Level 5) are not a function of the data, but a representation of aspects of student's proficiencies. Evidence from a student's performance updates our belief about her values for SMVs, which express what we know at a given point in play.

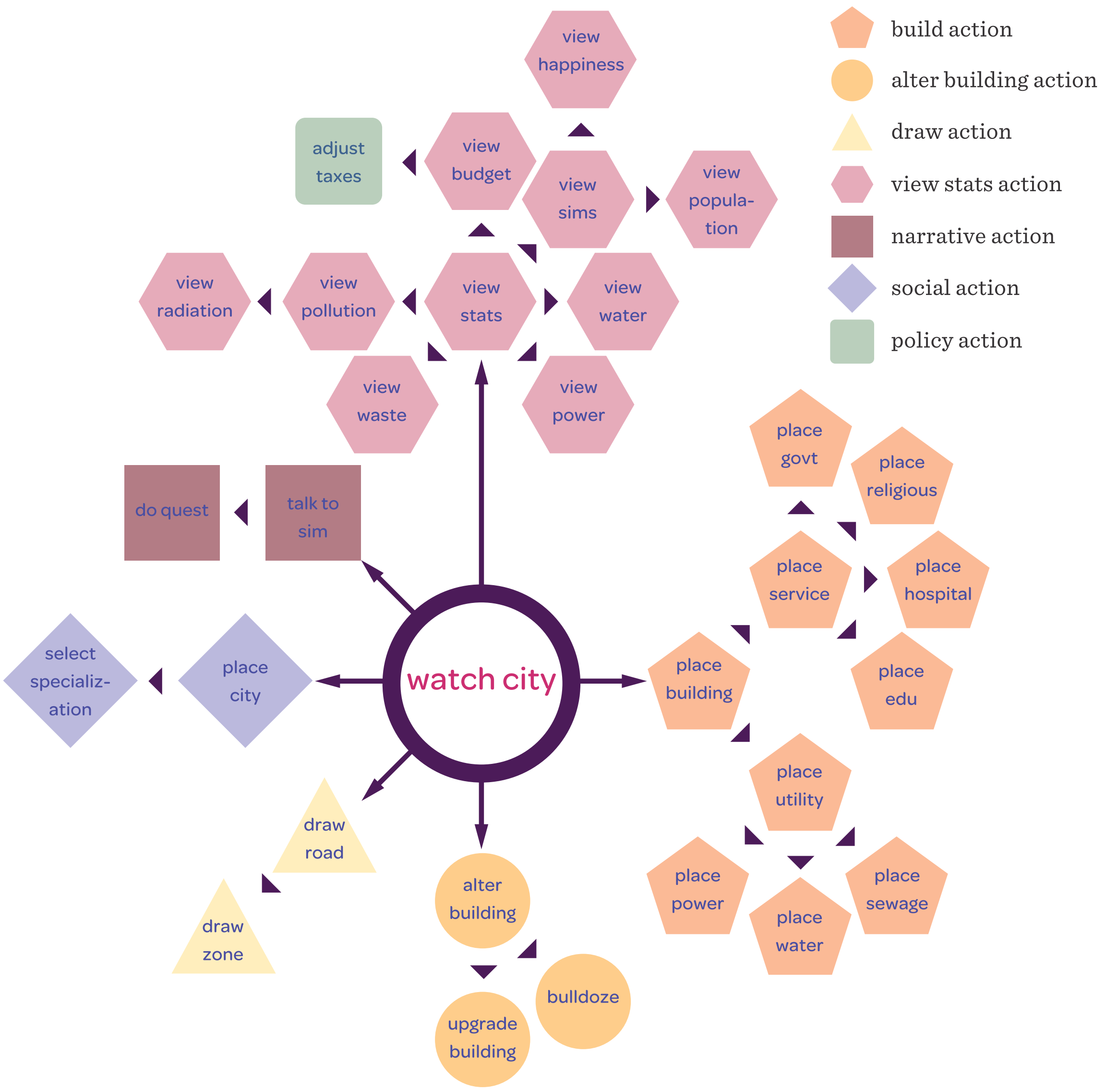
Assessment arguments culminate in claims about students, generally about either proficiency structures or propensities to act in certain ways. Claims may be strong and intended to apply widely outside the GBA, or they may be finely-grained, rapidly changing, and useful only for the next decision on feedback in the game. Moving up the levels in Table 6, reasoning is increasingly less contextualized. We can draw some lessons with experience from a simulation-based coached practice system the Air Force supported for learning to troubleshoot the hydraulics systems in the F-15 aircraft (Steinberg & Gitomer, 1996). Specific actions that fix faults in the hydraulic systems in the simulated F-15 in Hydrive are substantially different from the hands-on work on a real plane, but the intent was to make the system components, information environment, and action choices similar enough that by the time we reach Level 4, the underlying knowledge, skills, and strategies are effectively the same.

The first step is determining work products, or some captured form(s) from everything that has happened in a game that will be the captured—because we currently believe it holds, or suspect it may hold, evidence we need for one or more of the assessment purposes discussed previously. As the basic ECD models in Figure 8 made clear, this is just the first step along a chain of reasoning. Determining what characteristics of work products are nuggets of evidence will be next, and is addressed in the following chapter on evidence identification. We will see in this chapter, though, work products are defined and captured with an eye to what we want to make inferences about, and have designed the game situations and player affordances to evoke. In particular we will see next that there will be some work products we already know a lot about, because we have designed certain situations and actions expressly to provide us certain kinds of evidence. Other work products we know less about up front, and anticipate exploration to identify salient patterns -- perhaps followed by refinements to the game to better focus types of evidence we find.

Table 6:
Levels Involved in Reasoning for Performance in Game-Based Assessments

Level 1: Epiphenomena	*Epiphenomena*, or manifestations that arise from the game-playing experience: RGB values of pixels on a screen, mouse clicks at various times and place, acoustic waves, eye movements, etc.
Level 2: Grammar Structures	The various kinds of epiphenomena generally follow some rules of organization and combination we will need to understand if we are to use them. Screen shots may contain lines and edges, mouse clicks on a keyboard can produce words and sentences, mouse actions can be radio button selections or drag-and-drop to hot spots, and eye movements can signal fixations and jumps (saccades).
Level 3: Semantic Structures	Actions and products of certain kinds, following the grammatical rules, can have meaning with respect to the game: pull up help, bulldoze a power plant, ask a nonhuman game character a question. These actions can arise from different combinations of raw movements, and the same raw movements can have different interpretations in different game contexts. They do not constitute game play itself, but they are the essential building blocks from which play is fashioned. A design goal in serious games is for elements that have meaning in terms of play to also have meaning in terms of the domain. For example, dragging and dropping particular tokens to particular places to fill in a Punnett square has a semantic interpretation in transmission genetics, but it may also have an interpretation as a puzzle that must be solved to move to the next stage of play in crossing dragons to get offspring breath fire but can't fly.
Level 4: Pragmatic Structures	Why was a student carrying out various semantic actions; that is, do they represent an attempt to check the effect of a move, to cause a change in the game state to move toward a goal, to re-play a thorny situation? Sequences of semantically meaningful actions generally arise from reasons, motivated by a player's goals and their current understanding of the game state and the underlying knowledge and skills the GBA is meant to target.
Level 5: Proficiency Structures	These correspond to student model variables in ECD: aspects of students' knowledge, skills, propensities, strategy availabilities, and so on. These give rise to pragmatic sequences of semantic actions, traces of which are contained in work products.

Figure 11:
Verb map from SimCity™



A Classification of Work Products from Game Play

The three basic kinds of work products that can be captured in GBAs are predetermined work products, contingent work products, and log file data, i.e., a trace of players' actions at some level of detail.

Predetermined work products

The idea of predefined work products was introduced in Chapter 3 when we discussed assessment paradigms in GBA, specifically the second paradigm. Predefined work products can range from answers to multiple choice questions during a pause in game action to a naturally constructed but required plan, artifact, model, report, representation, etc. in the course of working toward a game goal.

Predefined work products are most like familiar assessment, in that the assessment designer has most control over the circumstances under which it will be produced, directives to the student, form of production, and foreknowledge of what features hold evidence about what aspects of proficiency. In ECD terms, Task Models and Evidence Models have been constructed, and instances of the tasks are with certainty effected by the Presentation Process through the logic of the state machine rules.

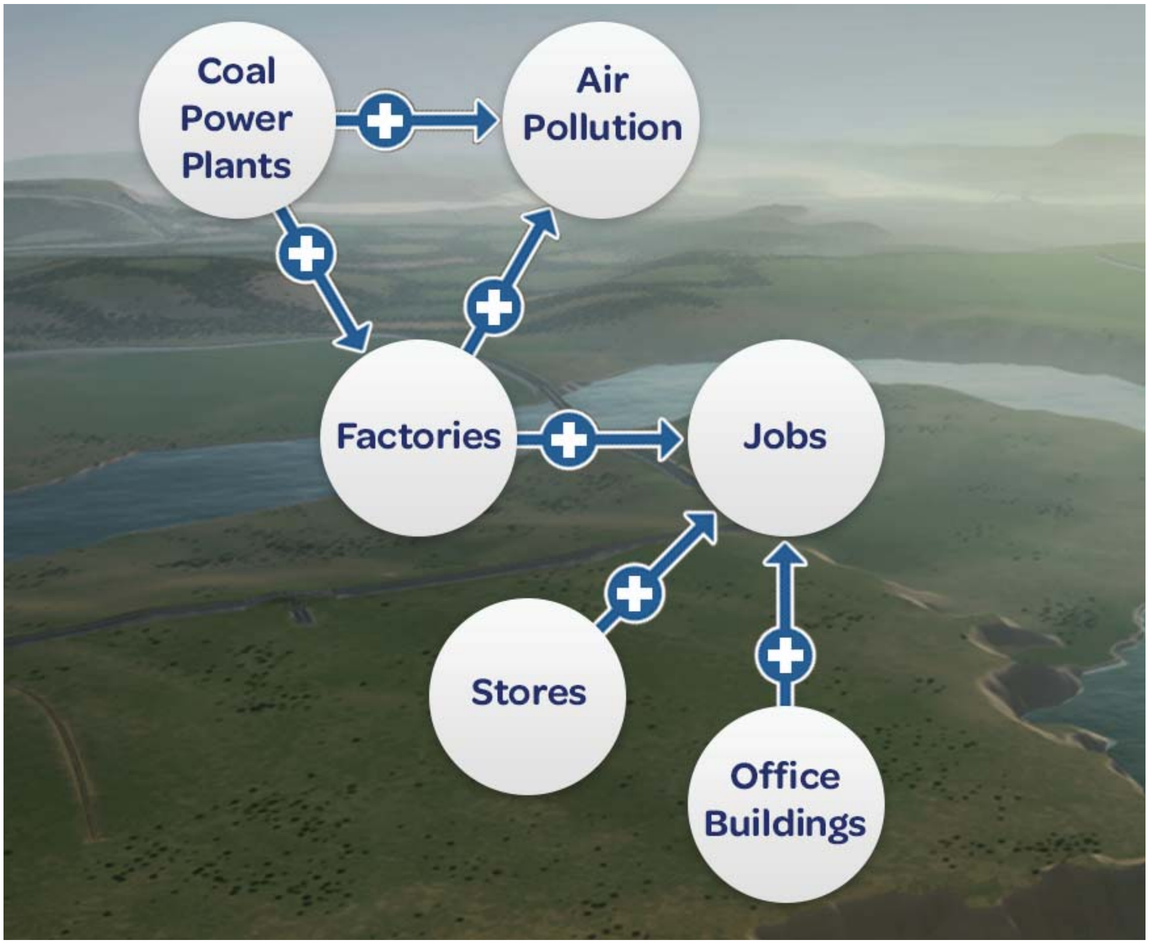
These features are all particularly important from the assessment perspective, but there can be a variety of ways a predeterminined work product arises in the course of play. A potential trade-off is this: Greater encapsulation of the activity as assessment can better focus the capture and evaluation of evidence, but it risks subverting the experience of game play. The most elegant solutions have tight assessment arguments (by controlling conditions they reduce alternative explanations for performance, minimize construct-irrelevant sources of variance) but feel integral to game play.

We mentioned previously some of the kinds of predetermined work products in Shaffer's Urban Science game: project reports, iterative planning diagrams, and a final zoning proposal. In Aspire, natural work products that arise as players work on computer networking contracts include final configurations of network designs for clients, the set up of networks, and the configurations with which either new networks are installed or faulty ones are fixed.

In Jackson City the system diagram is a predetermined work product. Both before and after tackling the challenge, a player takes a few minutes to sketch out their understanding of how key factors are related to each other and to the target output variables, pollution and jobs. Drag-and-drop of nodes, arrows, and labels makes for an activity that is fairly open-ended and constructive, yet yields an object with known and easy to parse components. Figure 12 is an example of a completed system diagram.

The system diagram, or causal loop diagram, details the relevant independent and the dependent variables as well as the relationships between them for the given scenario. As one source of insight into students' mental models of their cities, players complete their diagrams before and after the scenarios that require systems thinking. When they launch the systems diagramming application, players are provided with the key dependent variables to be explained. They then complete their concept maps by laying down new nodes for the independent variables and select variable names from a menu. Players are also responsible for showing relationships between the nodes by selecting and dragging one or more arrows between them. In addition they can change the direction of the relation between variables by switching the direction of their arrow. They control the valence of the relationship between two variables by selecting whether a given arrow signifies an 'increase' in the selected dependent variable or its 'decrease' through use of a specialized icon.

Figure 12:
Example of a Jackson City System Diagram



The number of independent variables presented in each node menu is constrained and so the set of possible diagrams is finite. Features of students' final products are captured by the assessment system and scored for the extent to which players' post-play diagrams capture multiple independent variables and for their accuracy in depicting the variables and relationships as presented during the students' gameplay. Information about students' processes to create the diagrams – changes in the nodes laid down in the workspace, amount of time taken for revisions and the amount of time given to the task as a whole, etc. – are also logged and become a potential source of additional evidence about students' competencies. We will return later to see how the observable variable MultivariateThinking score is determined from the system diagram, as well as an Accuracy score.

The post-play diagrams are one set of pre-determined work products. Another set of pre-determined work products – the end-state pollution, jobs and energy levels for players' cities – are more derivative of the players' game-play itself. These values are cataloged when players' solutions are submitted. The level of pollution, employment and energy are work products in the sense that student play within the simulation leads to their determination. In contrast to the systems diagramming tasks described above, the player's end state values are a natural consequence of students' play within the game and in that sense are more integral to it.

Contingent work products

Contingent work products arise from situations that lie between predetermined task situations presented to students and captures of unfiltered log file data. A contingent task is a recognized instance of a pre-identified, recurring, constellation of features that defines a class of evidence-bearing situations that can arise in play. When the defining conditions are satisfied, the presentation process is triggered to capture certain features of players' actions or ensuing game-state conditions (perhaps functions of them) because they are apt to contain evidence about certain aspects of proficiency. Such a segment of log file data—with respect to just the feature of the action that are pertinent and with any further transformations of it that may be needed—is a contingent work product.

An example is space-splitting situations in the Hydrive coached practice system for troubleshooting aircraft hydraulics systems (Gitomer, Steinberg, & Mislevy, 1995). Information available at a given point in a student's work, from the initial symptoms and his subsequent troubleshooting actions, defines an active path in a search space. Sometimes it is possible to carry out a test somewhere along the active path such that the results rule in or rule out a large portion of the problem space. This is called space-splitting, an effective troubleshooting strategy. Because different students attack the problem in a wide variety of ways, they work themselves into space-splitting opportunities at different times and different ways, and some students have many more than others. The Hydrive presentation process computes the effect of every move on a student's evolving active path, and when an agent detects the conditions of a space-splitting opportunity, it prepares to recognize the next set of moves as either space-splitting, serial elimination, remove-and-replace, redundant, or irrelevant. In other words, a "task" has been recognized, and a work product in the form of a set of actions that affect the active path in such a situation. It is a contingent work product because its appearance depends on the players' actions rather than being presented at a time and under circumstances determined wholly by the examiner.

The ECD task model structure can be extended to contingent work products. What remains the same are features that describe a situation and a description of the form of the work product to be captured. The difference is that the features of the situation are not used by task developers (or automated task generation systems) to create tasks to present; rather, they are conditions of ongoing situations in play that are monitored as dynamic task model variables. Certain configurations of their values at a given point in time signal the emergence of a contingent task, and prepare the Presentation Process to capture one or more work products of a particular kind. These in turn can be evaluated in terms of their salient features by Evidence Identification processes (as discussed in the next chapter). Note that the evidentiary of contingent work products, just as much as with predetermined work products, depends on the features of situations as well as features of players' actions.

Log file data

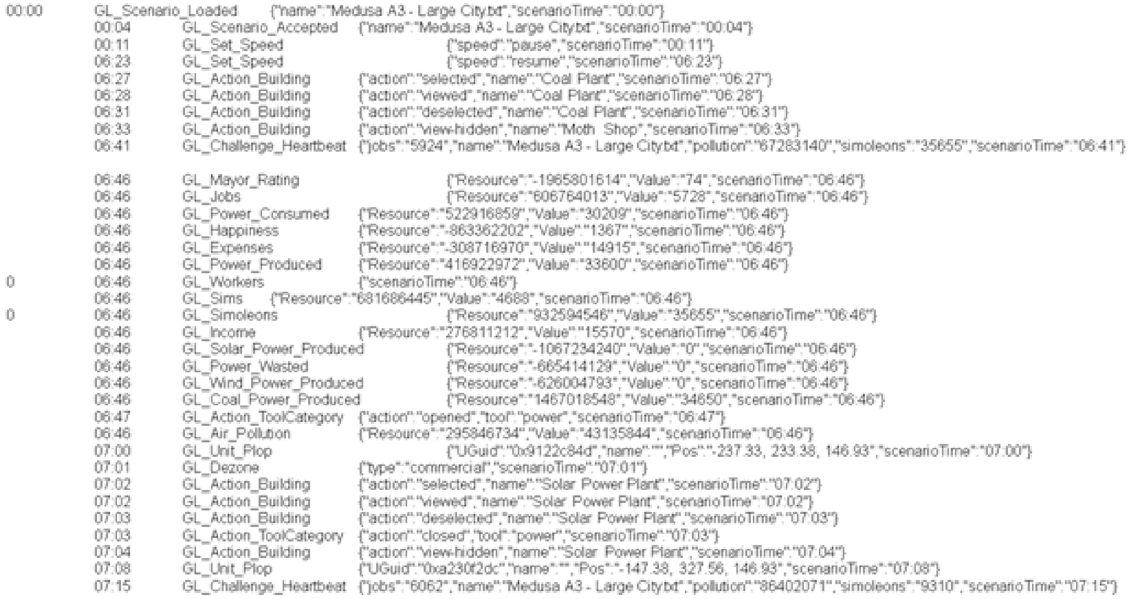
A log file is a work product that captures, at some level of detail and in some organized form, salient features of the status and activity in a game or simulation. Log files generally begin from unstructured records of player actions, or click-streams. These often take the format of a time stamp, an id, an indicator of location or place in the environment, an indicator of an action, and sometimes detail about that action. Two challenges exist with log file data: 1) deciding what information to capture and 2) identifying the elements of the log files that should be extracted and have scoring rules applied to them to create observables.

An initial impulse in deciding what elements of player action to capture in the log files is to collect everything (capturing data from play is called "telemetry" in the game industry; this term focuses on a role in a delivery system rather than a role in an evidence argument, as "work product" does; we will return to the connection presently). In reality there is often a tradeoff in developer time because the same people working on the game content are often the ones doing the coding of telemetry features (i.e., adding the code needed to capture each item of interest). In addition, in complex games the context and situations are changing by the millisecond and capturing every piece of information about play and game situation provides diminishing returns. The data will become more difficult to work with and the amount of information to be gained is likely reduced. Finally, although costs to store data are relatively small, storage is not free and the capacity needed to store vast quantities of raw information must be considered.

As a result, choices about what to capture must be made. Clearly any action already hypothesized to be an indicator of knowledge, skills, or attributes of interest must be included. Following that, it is helpful to capture video evidence of game play from individuals of different proficiency levels in a play testing step, and ensure that events of interest in the video play are captured in the log files. (Even when log files and video captures are separate work products, as might be used in play testing, it is possible to synchronize them in order to relate important episodes identified in the video with sequences of actions in the log file.)

Finally, there is consideration of how much information to capture about the game context in which actions are taken. A given action may have very different interpretations depending on where in the game it is taken. In the SimCity environment, for example, bulldozing a coal plant without having other power sources in place will result in an under-powered city with upset residents. The act of bulldozing can be effectual if pollution is high and there are alternate sources, but counterproductive if pollution is low and there are no alternate sources. In order to capture these dependencies on such paradata, a "heartbeat" (at regular time intervals or play-cycle intervals, capturing the value of salient variables in the environment) was inserted in the telemetry for capturing the state of pollution, power produced, and other variables in the city at very short regular intervals. Figure 13 is an example of a log file from Jackson City.

Figure 13:
A Segment of a Record of Player Actions From Jackson City



```
00:00      GL_Scenario_Loaded      {"name":"Medusa A3 - Large City.txt","scenarioTime":"00:00"}
           00:04      GL_Scenario_Accepted    {"name":"Medusa A3 - Large City.txt","scenarioTime":"00:04"}
           00:11      GL_Set_Speed                      {"speed":"pause","scenarioTime":"00:11"}
           06:23      GL_Set_Speed                      {"speed":"resume","scenarioTime":"06:23"}
           06:27      GL_Action_Building      {"action":"selected","name":"Coal Plant","scenarioTime":"06:27"}
           06:28      GL_Action_Building      {"action":"viewed","name":"Coal Plant","scenarioTime":"06:28"}
           06:31      GL_Action_Building      {"action":"deselected","name":"Coal Plant","scenarioTime":"06:31"}
           06:33      GL_Action_Building      {"action":"view-hidden","name":"Moth  Shop","scenarioTime":"06:33"}
           06:41      GL_Challenge_Heartbeat  {"jobs":"5924","name":"Medusa A3 - Large City.txt","pollution":"67283140","simoleons":"35655","scenarioTime":"06:41"}

           06:46      GL_Mayor_Rating                   {"Resource":"-1965801614","Value":"74","scenarioTime":"06:46"}
           06:46      GL_Jobs                           {"Resource":"606764013","Value":"5728","scenarioTime":"06:46"}
           06:46      GL_Power_Consumed       {"Resource":"522916859","Value":"30209","scenarioTime":"06:46"}
           06:46      GL_Happiness            {"Resource":"-863362202","Value":"1367","scenarioTime":"06:46"}
           06:46      GL_Expenses             {"Resource":"-308716970","Value":"14915","scenarioTime":"06:46"}
           06:46      GL_Power_Produced       {"Resource":"416922972","Value":"33600","scenarioTime":"06:46"}
0          06:46      GL_Workers              {"scenarioTime":"06:46"}
           06:46      GL_Sims     {"Resource":"681686445","Value":"4688","scenarioTime":"06:46"}
0          06:46      GL_Simoleons                      {"Resource":"932594546","Value":"35655","scenarioTime":"06:46"}
           06:46      GL_Income               {"Resource":"276811212","Value":"15570","scenarioTime":"06:46"}
           06:46      GL_Solar_Power_Produced           {"Resource":"-1067234240","Value":"0","scenarioTime":"06:46"}
           06:46      GL_Power_Wasted                   {"Resource":"-665414129","Value":"0","scenarioTime":"06:46"}
           06:46      GL_Wind_Power_Produced            {"Resource":"-626004793","Value":"0","scenarioTime":"06:46"}
           06:46      GL_Coal_Power_Produced            {"Resource":"1467018548","Value":"34650","scenarioTime":"06:46"}
           06:47      GL_Action_ToolCategory  {"action":"opened","tool":"power","scenarioTime":"06:47"}
           06:46      GL_Air_Pollution        {"Resource":"295846734","Value":"43135844","scenarioTime":"06:46"}
           07:00      GL_Unit_Plop                      {"UGuid":"0x9122c84d","name":"","Pos":"-237.33, 233.38, 146.93","scenarioTime":"07:00"}
           07:01      GL_Dezone               {"type":"commercial","scenarioTime":"07:01"}
           07:02      GL_Action_Building      {"action":"selected","name":"Solar Power Plant","scenarioTime":"07:02"}
           07:02      GL_Action_Building      {"action":"viewed","name":"Solar Power Plant","scenarioTime":"07:02"}
           07:03      GL_Action_Building      {"action":"deselected","name":"Solar Power Plant","scenarioTime":"07:03"}
           07:03      GL_Action_ToolCategory  {"action":"closed","tool":"power","scenarioTime":"07:03"}
           07:04      GL_Action_Building      {"action":"view-hidden","name":"Solar  Power Plant","scenarioTime":"07:04"}
           07:08      GL_Unit_Plop            {"UGuid":"0xa230f2dc","name":"","Pos":"-147.38, 327.56, 146.93","scenarioTime":"07:08"}
           07:15      GL_Challenge_Heartbeat  {"jobs":"6062","name":"Medusa A3 - Large City.txt","pollution":"86402071","simoleons":"9310","scenarioTime":"07:15"}
```

Further structuring can be useful, both to increase comparability across applications and to facilitate subsequent analysis. While games differ, the events in the games share some common features. In general, events can be categorized into system state events and player's activities. Both can be treated as a kind of "generalized action" that can be characterized by certain attributes and values. Among possible action attributes, the following are rather generic for all games: the player's id (PlayerID), the name of the action (ActionName), the time of the occurrence (ActionTime), who committed the action (ActionBy), to whom the action apply (ActionTo) to and the results of the action (ActionResult). Additional attributes can be defined as appropriate to particular games, states, and actions. Expressing the results in structured file formats such as XML produces files that look something like this:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<gamelog>
 <action>
  <PlayerID> bob </PlayerID>
  <ActionName> GL_Air_Pollution </ActionName>
  <ActionTime> 2/22/2013 3:57:54 PM </ActionTime>
  <ActionBy> system </ActionBy>
  <ActionTo> system </ActionTo>
```

```
   <ActionResult>
     {“Resource”:”295846734”,”Value”:”28439385”}
   </ActionResult>
  </action>
  <action>
   <PlayerID> bob </PlayerID>
   <ActionName> GL_Unit_Bulldoze </ActionName>
   <ActionTime> 2/22/2013 3:58:54 PM </ActionTime>
   <ActionBy> player </ActionBy>
   <ActionTo> Coal Plant </ActionTo>
   <ActionResult>
     {“UGuid”:”0xcc9cf003”,”name”:””,”Pos”:”52.92, -550.58, 142.28”}
   </ActionResult>
  </action>
</gamelog>
```

Once the log file is defined, the next task is to determine what information to extract for subsequent analyses. In most cases, there are some clear hypotheses about some particular actions that are related to the constructs of interest. Predetermined work products and known classes of contingent work products are examples. These can be extracted and scoring rules applied, as discussed in the next chapter. However, in many cases hypotheses are weak and new relationships between actions and knowledge, skills, and attributes can be uncovered with additional analysis. Exploratory data analysis and educational data mining techniques can be used to uncover patterns in the log files that are indicative of new or existing constructs. For example, Rupp et al. (2012) demonstrate how four different indicators in the log files were used to create a measure of the efficiency of a solution to a computer networking problem in a case where previously only correctness of the final solution was assessed. The indicators included time, number of commands, proportions of different types of commands, and amount of switching between computer devices. All of these elements were extracted from the log file.

Constrained-Response Work-Product Forms

Scalise and Gifford (2006) present a taxonomy of what they call “constrained response” task formats in computer-based testing. These can be adapted to serve jointly as game elements and work products in GBAs, both for predetermined work products and when suitable, contingent work products. They can produce focused and interpretable information for targeted capabilities, and as such are a basis of observable variables. Unlike their appearance in standard assessments, these work products need not be segregated and displayed as distinct “here is an assessment task” events. They can be, of course, and they often are in “gamified assessment” GBAs. But they can be integrated more seamlessly as part of play, within the game narrative, either as creating or completing some representation that makes progress toward a game goal (a predetermined or contingent work product) or as a natural sequence of actions (a contingent work product or a log file).

Figure 14 summarizes Scalise and Gifford's (2006) taxonomy; many examples appear in their article. Multiple-choice items are at the most constrained end of the degree-of-constraint dimension, while presentations and portfolios exemplify the least constrained forms. Game and simulation log files are in this furthest right unconstrained response column. Between them lie selection/identification, reordering/rearrangement, substitution/correction, completion, and construction types of tasks.

Designers can draw on any of these forms in a GBA. What this means for game designers is that as they develop features of situations, actions, and challenges to serve game-design purposes, they can have in mind ways of designing so that the same actions in those situations produce in one of these forms. In other words, there is a design bias toward eliciting actions that produce such forms, rather than in ways that are equivalent for game play but provide less comprehensible evidence. From the players' point of view, they are going through a holistic experience, driven by game-play considerations. From the assessors' point of view, the interactions capture evidence about players' thinking, and their capabilities more broadly construed, in a particularly interpretable form.

Joint game-assessment design patterns aren't necessarily easy to discover, but once they are, we would like to be able to re-use them. Rather than relying solely on designers' recollections and experiences, we can express them in shareable forms along the lines of software engineering design patterns (Gamma et al., 1994) or the assessment design patterns (Mislevy, Riconscente, & Rutstein, 2009) mentioned above: rationales, descriptions of when and how they can be used, examples, and, when feasible, sample code for implementation (e.g., for rendering forms, and algorithms for extracting and evaluating work products). Two examples:

- A generalized table format that needs to be filled out with drag-and-drop elements, that be used in a variety of GBAs for a student to express a provisional hypothesis and unlock laboratory tools to carry out experiments.

- A system diagramming tool, such as STELLA (Richmond & Peterson, 2001), which allows students to model a system with a palette of objects and connections that represent stocks, flows, feedback loops, etc., then run the model. The same underlying code can be used to present information as stimulus material, be a tool in investigations, and be used to create work products for a variety of subject domains and game contexts. It is minimally constrained, provides strong evidence about students' facility using systems concepts, and, because the work product is a file of objects and attributes, lends itself to automated scoring routines that examine its properties and check the results of its runs on standard test data.

This second example is in line with the systems diagrams students complete as a part of the Jackson City game. The Jackson City system diagramming activity can be seen as an example of task type 6c – the construct map - presented in Scalise and Gifford's taxonomy below. However, it is an example of a weak connection between the game play and assessment as it marks a break between the types of activities and strategies students are challenged with within the gameplay. Nevertheless, it also presents a good example of the sorts of trade-offs designers may be faced with as they consider how to meet the multiple demands that stem from the game and the assessment aspects of game based assessments.

The link between game play and activities associated with assessment is perhaps best preserved where students' systems thinking is assessed using evidence from their gameplay itself. This is the case in spite of the fact that the range of player actions within the game is fairly limited. While the SimCity play experience represented in Jackson City may seem quite free form, the set of possible actions and action-combinations that students can engage in is well constrained. Only a handful of discrete actions can actually be taken at any given time. What is generated via these limited categories of behaviors are large html files or vectors of telemetry describing the processes students engaged in to reach their final solutions. The resulting telemetry files describe each of the actions taken by the player, their timing and duration as well as the objects they were operating on. Viewed from this vantage point – the constraints imposed on the player and the data that results from their activities - the game-play itself is analogous to the figural constructed response tasks located in cell 6b of the Scalise and Gifford (2006) taxonomy.

Figure 14:
Intermediate Constraint Taxonomy for E-Learning Tasks (Scalise & Gifford, 2006)

Most Constrained → Least Constrained

Less Complex ↓ More Complex

Multiple Choice	Selection / Identification	Reordering / Rearrangement	Substitution / Correction	Completion	Construction	Presentation / Portfolio
True / False	Multiple True / False	Matching	Interlinear	Single Numerical Constructed	Open-Ended Multiple Choice	Project
Alternate Choice	Yes / No with Explanation	Categorizing	Sore-Finger	Short - Answer & Sentence Completion	Figural Constructed Response	Demonstration, Experiment, Performance
Conventional or Standard Multiple Choice	Multiple Answer	Ranking & Sequencing	Limited Figural Drawing	Cloze - Procedure	ConceptMap	Discussion Interview
Multiple Choice with New Media Distractors	Complex Multiple Choice	Assembling Proof	Bug / Fault Correction	Matrix Completion	Essay & Automated Editing	Diagnosis, Teaching

Figure 14
Based on "Table 1: Intermediate Constraint Taxonomy for E-Learning Assessment Questions and Tasks" from Scalise, K. & Gifford, B. (2006). Copyright 2006 by the Journal of Technology, Learning, and Assessment (ISSN 1540-2525).

The Role of Contextual Features in Work Products

The assessment argument shown in Figure 8 shows data concerning student's performances, data concerning the situations in which the performance takes place, and other information that may be known about the relation of the student to the situation. All three are used to make sense of students' actions in the ways the assessment is meant to address. The first of these, aspects of students' performances, is what people generally think of as "the data" in assessments, but all three play a role. The latter two are often tacit in familiar assessment. Designers and users can get away with leaving them implicit, because of the typical ways familiar assessments are designed and used, even though those assumptions are essential to the validity of the inferences being made. Standard practice and good instincts have just ensured that they are reasonably well satisfied. However, we find we must address them explicitly when we design new forms of assessment such as GBAs and simulations—in part because we have to write code to carry out inferences.

Background: Contextual features in familiar assessments

Most familiar assessments are built around predefined work products, such as test items, problems, and essay prompts. We can focus on just student responses and ensuing scores only because we are relying implicitly on the item writer to have determined just what should be in the situation (e.g., stimulus materials, tools, response choices for multiple choice tasks) and who will be assessed (e.g., language capabilities, background knowledge, knowing what is expected and how they will be scored). We attend only to the performances because we presume these things have all been set up appropriately. Exactly what an item writer has put into the features of a task, we assume they are tuned to evoke evidence we care about. The details and the rationale remain largely in the writers' head, trusted to her knowledge of the content area and the purpose of the test. They are often described in test specifications only to the general level of content areas and perhaps depth of knowledge (Webb, 1997) or Bloom taxonomy levels (Anderson, Krathwohl, & Bloom, 2005). Psychometric modeling of tasks' evidentiary characteristics is usually also modeled in terms of individual items.

More recent work in assessment design and psychometrics has brought task features to the foreground. There are several reasons for this, all of which hold advantages for interactive assessments such as GBA and simulations, in various ways for the different kinds of work products described above. The reasons that task features have become prominent in standard assessment include the following:

- Explicit connection to research on domain structure and learning in domains (Embertson, 1998).
- More explicit backing for validity arguments, in terms of "construct representation"; i.e., why actions in task situations should provide evidence about targeted aspects of students' capabilities (a consequence of the preceding reason).
- Better connection to the forms of psychometric models and operating characteristics of tasks (Adams, Wilson, & Wang, 1997).
- Automated task construction (Gierl & Haladyna, 2012).
- Adaptive assembly of assessments from tasks with known evidentiary properties; i.e., what aspects of proficiency are evidenced, at what levels, and how much evidence is obtained (Almond & Mislevy, 1999).

Contextual features and predefined work products in GBA

In GBAs, predetermined work products are most like tasks on familiar assessments. Even here, there are significant advantages for making salient situation features explicit. Identifying features of task situations that tend to elicit certain aspects of students' proficiency helps keep the design of a GBA on target with respect to both learning and assessment, when cross-disciplinary design challenges are simultaneous. This is especially important in GBA design because more than one designer is usually involved and their expertise needs to be used jointly. It no longer suffices to count on an item writer's personal knowledge to craft the artifact, because it also must also meet constraints she is not an expert in. Explicit representations help people with different knowledge work together, especially when the representations are structured around key relationships that may not be visible on the surface (Collins & Ferguson, 1993). The preceding section noted the value of a taxonomy of work-product forms at hand to help game designers embed them in game play whenever it was natural. Having explicit task model variables similarly helps them incorporate features into game situations that are particularly useful to evoking thinking through targeted proficiencies.

A further advantage for explicitly encoding some information about task situations for predefined work products is that task model variables bear information about task characteristics such as difficulty and what aspects of proficiency are evidenced. This means it is not necessary to collect as much data to start up psychometric modeling in early use, and in fact approximations of psychometric model parameters based solely on task features can suffice in low stakes applications (even high stakes ones when the relationship is strong enough, such as in the British Army Recruitment Battery (Irvine, in press)). The section on structured psychometric models in Chapter 10 will discuss these ideas further.

Contextual features and contingent work products in GBA

As described in the previous section, contingent work products arise when certain recurring configurations of situations that arise in the course of game play signal an evidence-bearing situation in a known way. The example was an opportunity to do space-splitting in troubleshooting in Hydrive.

Monitoring contextual features is required to recognize these situations. Whatever contextual features are required to signal such a configuration need to be monitored in phases of the game where they can arise. In ECD terminology, these are called run-time or dynamic task model variables. The variable characterizes some aspect of a situation that can take different values—present or absent, how much, how many times so far, whether it is day or night, how many other players are in the same room, what is the current level of pollution and number of coal plants, etc.

Some of these contextual features may already be calculated and posted in an ongoing game-state table. If so, they can be monitored by an agent for each such continent task. It may be that it is functions of already-calculated contextual variables are required, so they can be computed and put in the state table, or alternatively calculated by the agent. The latter takes a little more computing, but does not expand the game state table and better separates the assessment routines so they can be modified or added more independently. Further, detecting the conditions for recognizing the opportunity for a contingent work product may additionally require information about the player: Exactly the same features of a room in a game can be an evidence-capturing opportunity for a player on his first visit but not subsequent visits, for example, or only if he has already shut down enough coal plants.

As with predetermined work products, these contextual features can be used by game designers as they craft elements of game play: other things being equal, they should structure game situations that lend themselves to provoke contingent work products.

Also as with predetermined work products, the values of dynamic task model variables contain information about what aspects of proficiency are being evidenced and the nature of that evidence (via parameters of psychometric models). What is different now is that contingent tasks are one-off; we cannot count on having multiple identical presentations of them, as are needed to calibrate tasks in standard assessments (i.e., to estimate their parameters under the psychometric model). It is only the values of the task model variables that are available. In low-stakes applications, using expert opinion to build a function that maps values of task model variables to values of task parameters may suffice. When evidence is available from many (distinct) instances of a given class of contingent work products, statistical methods are available to refine these approximations (Glas & van der Linden, 2003).

The dog that didn't bark in the night

With predetermined work products, the key contextual features are designed into situations that are expressly presented to the student. In contrast, capturing contingent work products is only possible when we recognize that triggering features of situations have occurred as a result of game play. Instructional designers and educational data miners have adopted the term Paradata, the contextual data that accompany response data mentioned previously, proves important to interpreting actions and making decisions.

Using contextual data to distinguish contingent work products makes it possible to not only look at actions actually taken, but to examine actions in terms of situations where classes of actions could be taken—including no action taken at all. The "curious incident in the nighttime" in the Sherlock Holmes story "Silver Blaze" was that the dog did not bark—a clue that the thief was his master. In Hydrive, for example, doing 12 space-splitting actions in a given problem is generally better than doing 6. But there is stronger evidence if we know the first student did 12 space-splitting actions in 20 situations when they were possible, while the second did 6 out of 7. The latter is now seen to suggest a higher level of proficiency. Additional evidence accrues by seeing which of the 8 of 20 situations where the first student could have space-split but didn't were serial elimination, remove-and-replace, redundant, and irrelevant actions.

In statistical terms, this is a move from modeling Baconian events, or "things that happened," to Pascalian events, or "things that happened in a defined space of things that could have happened" (Mislevy & Gitomer, 1996). More powerful statistical tools—in our case, more powerful psychometric models—are available to capture, characterize, and manage evidence, in more complicated and subtle situations (Schum, 1994).

Contextual features and log-file work products in GBA

Log files are attractive because they can contain a great deal of data, hence potentially a great deal of evidence, garnered in relatively unconstrained situations. Even a small proportion of evidence in a massive amount of data might result in a substantial amount of evidence, if we can just identify patterns that bear evidence. Automating the discovery, then the routine characterization, of "feature detectors" is a central problem in machine learning (Gong, Ng, & Sherrah, 2002). We are particularly interested in it in simulation and game-based assessment, as feature detectors are the basis of what become observable variables in ECD terms (Gobert, Sao Pedro, Baker, Toto, & Montalvo, 2012). We will say more in the section about evidence identification about how this problem can be approached, but address here a more specific question: "How much and what kind of paradata do we need to include about the ongoing game situation to exploit this potential?"

In general, actions hold more meaning, hence greater evidentiary value, if we know more about the situation in which they are carried out. (This is like the difference between a dictionary definition of a word and its meaning in a context of use; compare "Somebody spilled the coffee; get a broom" and "somebody spilled the coffee; get a mop." (Gee, 2013)) On the other hand, the demands of telemetry, data storage, and analysis impose limits on just how much paradata can be carried along in analysis. Initial investigations—data mining—might be carried out offline with more contextual data to discover feature detectors, then only those found useful will need to be routinely calculated and monitored in real-time use.

It is interesting to note differences in the design space of contextual data use for various assessment purposes:

- Traditional test items, written only to broad test specifications, contain very specific context but the details are not made explicit. Such specifications are chosen to be maximally useful, but their usefulness and the evidentiary value of the items is tied to those specific items.

- Psychometric models that require encoding key task features operate at a finer level of detail, and as mentioned, lend themselves to "borrowing information" to make inferences to other items with various configurations of the same features. That is, knowing some of the key features of items that define their local context makes it possible to know a lot about what aspects of knowledge and skill they will require, and how hard they will be. Sometime we can even use item-level contextual features to approximate the evidentiary characteristics of new items as they are recognized or are generated on the fly (Bejar, Lawless, Morley, Wagner, Bennett, & Revuelta, 2003).

- Discourse analysis and conversation analysis, as they are carried out by applied linguists and social scientists, address language use in light of deep and detailed analysis of context. Not only are contextual features included as they might be characterized objectively, but as they might be characterized by individuals, in light of their unique experiences, resulting in different contextual information as interpreted for different people (Gee, 2013).

- In contrast, some automated scoring schemes with massive data operate successfully with very little contextual information. Latent semantic analysis (LSA) of huge text corpora can perform quite accurately using only co-occurrences of words, for tasks document retrieval, essay scoring, and even taking vocabulary tests (Landauer, Foltz, & Laham, 1998). The only contextual information used is that the words appeared somewhere in a given corpus of texts. This feature is not trivial, though; LSA representations of texts can differ for, say, medical texts and legal texts. Perhaps the earliest example of this approach carried out with modern statistical methods was establishing the authorship of disputed Federalist Papers (Mosteller & Wallace, 1963).

In determining what context features to capture in-game, simulation and game designers need to consider both what features should be captured and how often. In the design of SimCityEDU games like Jackson City, a "heartbeat" was established in the log files. In SimCity, time is constantly passing, and the simulation engine is constantly running. Even if a player does nothing in the city, the context will change as the elements of a city would change without intervention. For example, if there are a lot of coal plants, the air in the city will continue to become polluted; the pollution numbers will rise over time. People may begin leaving the city and population numbers will fall. All of a player's actions occur at a specific point in time with associated levels of each of these measures of the health of a city. Their actions may or may not be a reaction to these measures (which players can monitor). It is important to know if a player is bulldozing power plants in a city where there are dangerous levels of pollution or in a city where the pollution is at an acceptable level. Since these numbers are constantly changing, it was not feasible to capture each change. Instead, a heartbeat was created so that every 30 seconds of game time, the levels of the primary measures in the city were captured and stored. With this information, it was determined that sufficient context could be established for most actions in the game. Variables such as "bulldoze in low pollution scenario" and "bulldoze in high pollution scenario" can now be created and used in analyses to detect patterns in game play.

Context can include not just variables from within the digital environment, but also information about the setting and events surrounding the use of the digital tool. In an analysis of a simulation-based assessment of computer networking skills, Rupp et al (2012) examined a tool that was designed as a formative assessment. The simulation-based tasks were designed to prepare students for a skills assessment using real networking equipment. However, examination of the data revealed entire classes for whom the simulation-based assessment was the final skills assessment. Separating the students who had completed the assignment in a summative context from those who completed it in a formative context revealed differences in the fit of the scoring models. This finding underscores the importance of considering many aspects of the real and virtual environments.

There are not clear rules regarding how many and what types of context variables are important. Thinking forward to the types of analysis to be conducted can help, but watching individuals work through a simulation or play through a game with an accompanying think aloud protocol can be most instructive. When watching players play and describe their actions, it is often quite clear how to interpret their actions in the game. The question then becomes whether there are player and game events that can be captured that convey, even if imperfectly, those influences without the analyst actually observing students play and hearing them talk. These events then become the variables it is important to capture in work products and be able to identify automatically

.

Telemetry and the Four-Process Assessment Architecture

Telemetry is a general term used in technology fields to mean automated collection of data from a remote location and transmission to other receiving equipment. In gaming, it is used to mean the collection of information at the point of player interaction with the game and its transmission back to servers for collection, storage, and analysis. In ECD terms, telemetry data is material that can, perhaps after some processing, yield a work product. In other words, telemetry data has been gathered from a player's actions and associated game states, as the result of the interaction of a player with an activity. It is raw material from which work products are identified. Sometimes little or no processing is needed, as when choices to discrete multiple-choice items would be sent as telemetry data. Each response in an adaptive assessment, or a vector of them in a fixed test, would constitute work products. Telemetry data in the form of mouse clicks and pointer hovers might require interpretation and further augmentation with contextual features to produce a log file work products. Rules can then be applied to such work products to extract and characterize information, to create observables, in evidence identification processes.

Captured data can also be used immediately, in tighter cycles of evidence identification, evidence accumulation, and activity selection as play occurs. When certain scoring rules are known, they can be applied to logged data immediately and trigger feedback on the activity, as well as hints and support, provided in-game. For example, using the heartbeat described above, there might be a rule that indicates when pollution climbs over a given number, the player should be given a warning that pollution is too high. The player is interacting with the city, creating a work product that is the log file. The pollution level is being extracted out of that log and continually compared to the threshold number in the evidence identification process. If it is below the threshold, the score is 0 and if it is above the score is 1. This score is accumulated by the simple rule of addition, so the score goes up by 1 for every 30 seconds that the city is over the pollution threshold. This profile is then used to determine whether and when to trigger the pollution warning (conceptually, an "event selection" in the four process model; operationally, an action triggered by a rule in the finite state machine based on the value of this game-state variable).

In the early stages of game development, the ways in which evidence can be identified from the log files constructed from telemetry data are often not clear. In this case, the data sent to the central servers from many players is analyzed with exploratory data analysis and educational data mining techniques to look for relationships in the data, identify classification systems for players based on patterns in their play, and create various models that can then be used to improve the scoring rules (or suggest modifications to game features and player affordances).

For example, in a Jackson City scenario in which the final levels of the city's power and pollution might indicate whether a player understood that causes have multiple effects (an element of systems thinking), data mining might reveal that the sequence in which players build and bulldoze buildings is also related to this understanding. That is, if players bulldoze coal power sources prior to building cleaner power, it may be an indication that they are focusing only on a single cause and effect relationship (coal to pollution), rather than its multiple impacts (providing power and pollution). Occurrences of this sequence could then be extracted from each player's data and the count included as a new observable—another piece of evidence about systems-thinking proficiency.

Decisions about what data are used locally and what is sent to remote locations are based on tradeoffs among technical and logistical considerations. For example, the speed and openness of the networks on which a game is likely to be deployed may influence how often and what size data files can be passed from a local site to a remote server location. In a cloud-based game, data passes back and forth from the cloud to the device throughout play with no information processed locally. In a massive multiplayer online game this is a requirement in order for the interactions of all the players to be apparent to each other. Alternately, some games are completely local, passing no data to a central location. In general, uses of games in assessment contexts require at least some communication with a central location, but this frequency and size should be determined with the implementation context carefully considered.

Identifying Evidence

This section addresses how to identify and characterize features of work products that hold evidence about what students know and can do. In general, they concern semantic and pragmatic aspects of performance, expressed in terms of values of observable variables (perhaps after multiple steps of processing). Evidence identification is critical in several senses: It lies at the center of the bridge from tasks to inferences about students' capabilities in the ECD models (Figure 8). It is an essential link in the reasoning chain of the assessment aspects of any game-based or simulation-based assessment. And it is a leading edge of assessment research as technology enables us to capture ever richer and more complex performances—and we need to be able to make sense of them (Rupp, Nugent, & Nelson, 2012).

Evidence Identification with Predetermined Work Products

Evidence identification is more straightforward in pre-determined work products, and particularly so for those toward the more constrained forms of the Scale-Gifford taxonomy (Figure 14). Regarding pre-determination, as noted above at least some of the key contextual information for interpreting actions is known, and the permissible actions have been constrained to be semantically and/or pragmatically meaningful. Levels 2-4 in Table 6 will have thus been handled up front.

Just because meaningful actions are captured in the work product, however, does not necessarily mean it is easy to identify and characterize the patterns and features of those actions that constitute evidence.

It is easy toward the constrained side of the hierarchy because the actions have been constrained to produce only semantically meaningful patterns. For example, discerning evidence in familiar encapsulated multiple-choice tasks is easy, because all of the design work has been up front. The features of the situations have been crafted to evoke the targeted thinking among the testing population, and the choices provide clues about that thinking. The knowledge and skill can be anywhere from simple recognition of definitions to complex reasoning and evaluation, but the form of the work and paired evidence identification procedure is straightforward: Did the student choose the keyed option? (The value of the resulting observable variable still provides only uncertain evidence about student model variables (SMVs), and it might not even be very much; the point here is that determining the value of the observable variable (OV) itself is easy and reliable.)

As we move to the less constrained work-product forms, even though we know where to look and what it is supposed to bear evidence about, the patterns that constitute evidence can be less clear cut. Many scoring algorithms have been proposed over the years for concept maps, for example (Ruiz-Primo & Shavelson, 1996). Automated scoring of essays can involve quite sophisticated uses of natural language processing (NLP) techniques to identify linguistically relevant features, then a further step using methods such as logistic regression or neural networks to capture subtle patterns that correlate with semantic meaning—e.g., matching human raters' holistic scores or evaluations of style and mechanics (Deane, 2006). A fair amount of tuning, exploration, and data mining can be required to extract meaning even in predetermined work products with more open responses.

Jackson City Example: Observable Variables from the System Diagram

Figure 12 showed an example of the Jackson City diagram that players draw to express their understanding of the relationships among factors in the pollutions and jobs problem. The final collection of directed links and positive or negative associations between pairs of factors that a player places on the diagram is a work product. In prototyping, two observable variables were generated from this work product: MultivariateThinking (Table 4) and Accuracy.

The evidence-identification rules that produce MultivariateThinking are effected in a series of three steps. From all the actual movements, mouse clicks, timings, and possible placements and removals of links that produced the diagram (not to mention possible sweat and tears), all that is saved is the ordered triples that represent which two factors have been linked, in which direction, and positive or negative influence. This information is saved in a matrix, with all possible off-diagonals entries either a 0 for no link, 1 for a positive relationship, and 2 for a negative relationship.

Primary features extracted from this matrix include a categorization of the resulting set (Undifferentiated, Some Organization, or Multilevel Structure), counts of factors linked to both jobs and pollution, use of all factors, and counts of links that are correct and that are incorrect in terms of the underlying system relationships. Final rules applied to the set of derived features categorize each solution into the levels shown in Table 4. Another set of rules is run on the counts of correct and incorrect links to the dependent variables jobs and pollution to produce the Accuracy score, with levels 1) Neither jobs nor pollution direct links are accurate; 2) Links to just one dependent variable are accurate; and 3) Links to both dependent variables are accurate.

Once the work product has been defined, it is possible to revise the definitions of observable variables or add new ones. With a data set of many players' diagrams and game play, we can empirically search for functions of features that correlate well with other indicators of systems thinking, such as degree of success for the final state of the game. When work products are complicated, even if they are predetermined, we can use data mining techniques such as the ones discussed in the following sections to discover additional evidence (and in the process, gain insights to improve the design of situations and affordances).

Evidence Identification with Contingent Work Products and Log Files

Much of the excitement about game-based assessment is about being able to capture fine-grained data about player activity. The promise is that this data will help us understand the processes that players use to solve problems, not just their final products. It is argued that there is great potential for generating new insights regarding complex knowledge, skills and attributes.

However, the potential of games as assessment tools can be met only if methods for making sense of stream or trace data (in familiar terms, "scoring it" [6]) can be developed in evidentiarily sound and computationally feasible ways. Traditional psychometric models have commonly been focused on point-in-time models that overlook variation in activity over time (especially at the micro level). New interactive digital experiences such as on-line learning environments and games, however, elevate both the availability and importance of understanding student temporal micro-patterns, which can reflect variation in strategy or evolving psychological states. While the richness of the data holds promise for making important inferences, few standard methods for scoring and analysis exist.

A primary challenge in fulfilling the potential of log files for making inferences about students thus lies in evidence identification. We have tasks that are often open and multifaceted, in which learners can interact with the digital environment in a number of ways, choosing various paths through the game environment. What are the important features of a work product and how do we apply scoring rules? Log files present many types of data, including sequences, frequencies, and duration of actions. Potential evidence for each construct must be gleaned from the masses of potential data available. We must determine how to turn this evidence into values of observable variables. In traditional multiple choice tests, scoring is quickly accomplished by evaluating each response as correct or incorrect. When assessing new constructs with new forms of data, the simple notion of "correctness" may no longer be good enough.

Currently, the problem of evidence identification in log files is often tackled by combining a priori hypotheses about the relationships between observables and constructs with exploratory data analysis and data mining (Mislevy, Behrens, DiCerbo, & Levy, 2012). For example, Rupp et al. (2012) describe this process for an activity involving configuration of a computer network. The researchers initially ran confirmatory model fit analyses to examine the relationships between the observables and the skills they are hypothesized to measure. They then conducted exploratory analyses to examine log files consisting of time stamped commands that students entered to configure computer networking devices on a simulation-based assessment. They identified features including the number of commands used to configure the network, the total time taken, and the number of times in the log that students switched between networking devices, as evidence that could be combined into a measure of efficiency. This combination was arrived at through a series of analyses using multiple statistical traditions including tagging commands as done in the Natural Language Processing literature, visualization of patterns with sociograms, and principal components analysis. Note that

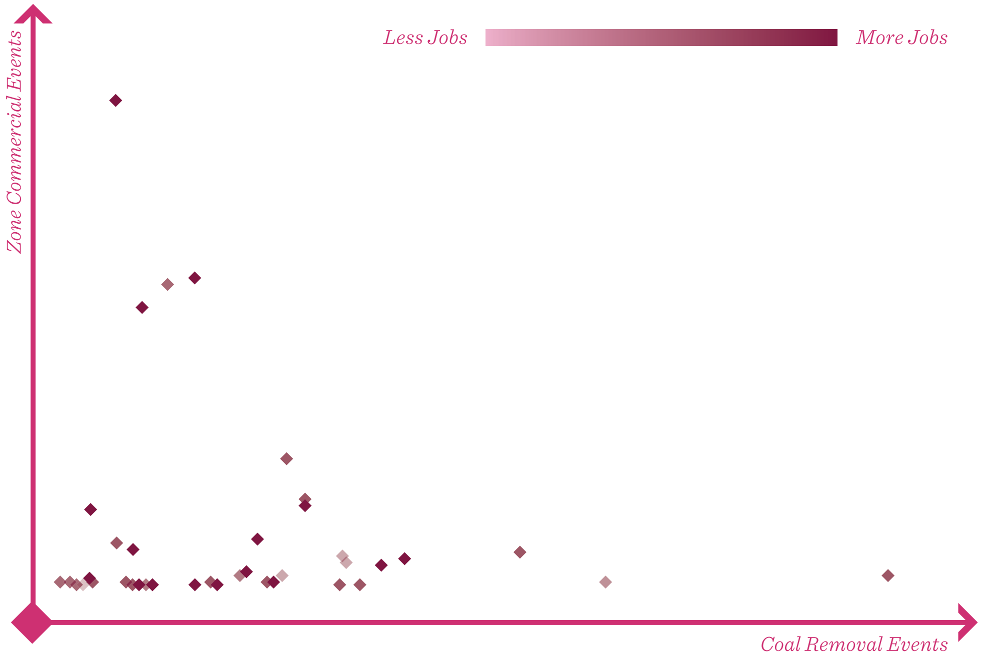
[6] We sometimes use the term "scoring" at times because it is familiar, but familiarity is a disadvantage when it constrains thinking about what to look for, how to characterize it, and how to use it (Behrens et al., 2012). The less familiar terminology of ECD is more useful because it situates thinking in the realm of evidentiary argument more broadly, and allows us to talk in ways that apply to familiar assessment but also, in a rigorous way, to the more complicated challenges that arise in unfamiliar forms of assessment such as GBAs.

none of these features was scored “correct” or “incorrect,” and their combination provided evidence for inference about students’ capabilities, apart from the overall correctness of their performance.

Exploratory Data Analysis (EDA) is a conceptual framework aimed at providing insight into data, and to encourage understanding probabilistic and non-probabilistic models in a way that guards against erroneous conclusions (Behrens, DiCerbo, Levy, & Yel, 2012). EDA provides a conceptual framework and set of heuristics for pattern discovery, hypothesis generation, and assessment of rough confirmation in interactive agile cycles. John Tukey championed EDA in the 1960s; the approach has expanded rapidly with burgeoning digital technologies and massive data bases. Using a variety of tools, EDA encourages the exploration of the patterns in data and the potential explanations for those patterns. Then, the most promising hypotheses can be extracted for further testing using more confirmatory methods. The four key heuristics of EDA are: Revelation (through visualization), Re-expression (re-scaling), Residuals (iterative model building and revision), and Resistance (statistical care against unusual observations).

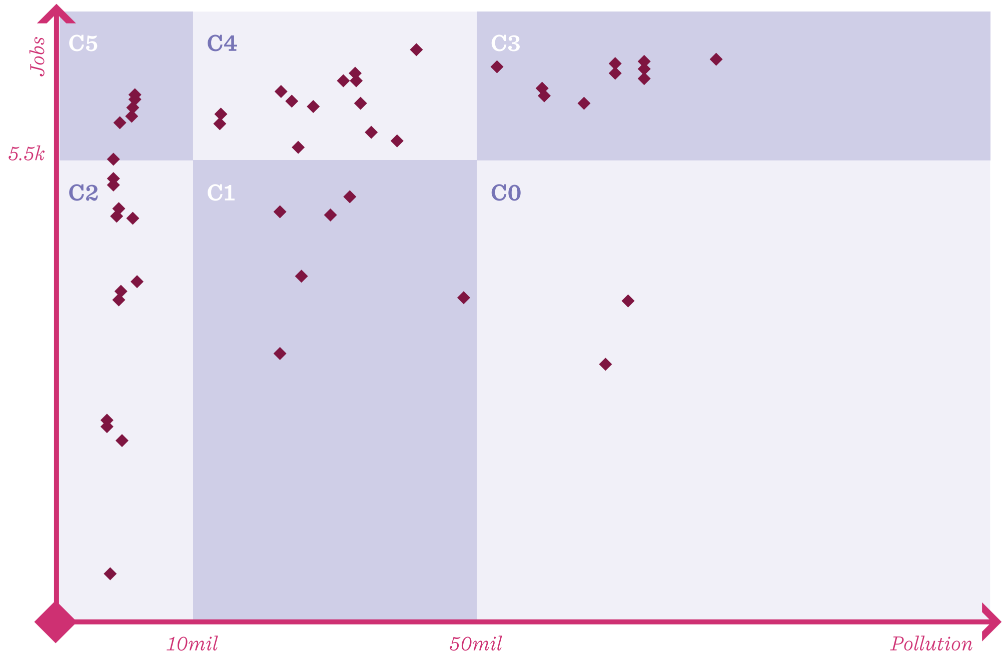
EDA is useful in developing evidence identification from log files because often we start with only vague hypotheses about how features in log files might relate to each other and the constructs of interest. Data visualization can assist in developing hypotheses that may inform both assessment modeling and game design. Simple scatterplots, for example, can reveal patterns. In an early version of the Jackson City challenge, success was defined simply as having low pollution and high numbers of jobs in the city. Theoretically, eliminating coal plants and then zoning areas of the city as commercial should result in these outcomes. The number of instances of coal removal and zoning can be extracted from the logs. After an initial round of play testing, the scatterplot in Figure 15 was used to examine the relationship between coal removal, zoning, and jobs numbers. Each player’s indicator is colored according to the final number of jobs. A quick review of the plot revealed that many players were able to achieve high levels of jobs doing little or no zoning, an undesirable feature of a solution that would satisfactory in terms of the substance of the problem. This finding led to a reexamination of the in-game algorithms that created the jobs numbers, and to extensive redesign of the tasks.

Figure 15:
Scatterplot of Events in SimCity Colored by Number of Jobs in the City



Further refinement of the game situation led to the end-state values of pollution and jobs of playtesters shown in Figure 16. This is the information that led to the (provisional) definition of the Observable Variable called JobsPollutionEndstate described above in Table 5. The work product being addressed is the final configuration of the game. The features of this work product that are involved in this OV are endstate pollution and endstate jobs. The observable variable is defined by the regions labeled C1 through C5 in the following figure:

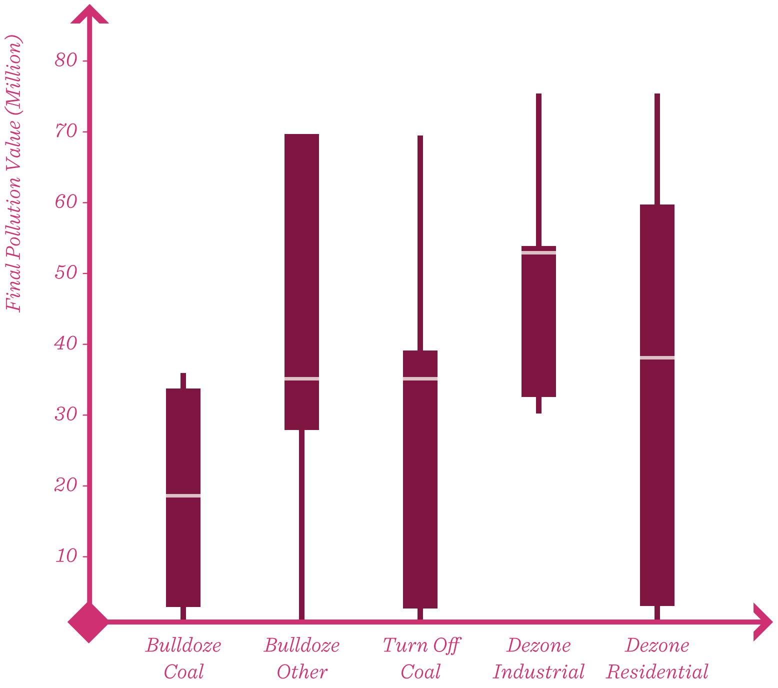
Figure 16:
Ending States of Pollution and Jobs



Level	Cell	Evaluation Rule
5	C5	Pollution < 10mil Jobs > 5.5k
4	C4	Pollution > 10mil and < 50mil Jobs > 5.5k
3	C3	Pollution > 50mil Jobs > 5.5k
3	C2	Pollution < 10mil Jobs < 5.5k
2	C1	Pollution > 10mil and < 50mil Jobs < 5.5k
1	C0	Pollution > 50mil Jobs < 5.5k

In a second example from the early Jackson City analysis, the relationship between the first action taken and final levels of pollution was examined. As shown in Figure 17, players whose first action was to bulldoze coal plants or dezone residential areas both ended up with low median final pollution scores. These results suggest the importance of the first actions to ultimate success in the scenario. Bulldozing other buildings, for example, may suggest a misunderstanding of the causes of pollution in the city or a lack of understanding of the goals of the game. The visualization served as an impetus to begin forming hypotheses about the relationship between early actions in the game and both causal understanding and game comprehension.

Figure 17:
Box Plots of the Final Pollution Calues by First Action Taken



The techniques of EDA in general require a great deal of human intervention, with the analyst acting as a detective, uncovering patterns in the data. However, growing data sets and computing power have also led to the rise of Educational Data Mining (EDM). Educational data mining is the process of extracting patterns from large data sets to provide insights into instructional practices and student learning (Romero et al., 2011). It can often be employed for exactly the tasks of evidence identification: feature extraction based on patterns in data. Kerr and Chung (2012) conducted exploratory cluster analyses to identify salient features of student performance in an educational video game targeting rational number addition. DiCerbo & Kidwai (2013) used Classification and Regression Tree (CART) analysis to build a detector of whether game players were pursuing a goal of completing quests (as opposed to other potential goals) in a game environment.

The CART process is an example of the types of machine learning techniques that can be applied to log files to identify features and rules. These techniques require a categorical outcome variable, a set of potential predictor variables, a set of "learning" data by which to establish rules for classification, and a set of "test" data by which to validate those rules. Researchers develop a sample with known values on the construct of interest (for example, the player's goal), then attempt to identify elements of the log files that can predict each individual's status. A set of hypothesized features—elements of play recorded in the log file—are identified that an automated detector could use to classify an individual.

The process of creating decision trees begins with the attempt to create classification rules until the data has been categorized as close to perfectly as possible; however, this can result in overfit to the training data. The software then tries to "relax" these rules, in a process called "pruning" to balance accuracy and flexibility to new data. A variety of pruning algorithms can be used to try to find the easiest, most interpretable tree.

The result of the analysis is a decision tree, or graphical representation of the series of rules for classifying cases. The nodes of the tree denote features, the branches between the nodes give the rules for the values of that feature to be used for classification, and the end nodes of each branch give the final classification of a case. A given individual's classification is then the value of an observable variable procedurally defined by this decision tree. The CART process is just one of many types of machine learning, but provides an example of how these algorithms can be used to identify important features and scoring rules in log files.

The "feature detector" approach mentioned in Chapter 8 is an example of an approach that iterates between humans and machine learning. In an open-ended activity space such as a GBA or simulation task, human experts look for instances of important action sequences, which are classified as instances of higher-level, semantically interesting, features. These tagged sequences are used as prediction targets by functions of lower-level features that can be automatically detected without human intervention. Specific approaches for this modeling step include regression and logistic regression (Margolis & Clauser, 2006), neural network modeling (Stevens & Casillas, 2006), and Bayes nets (Scalise, 2013). In this way, we obtain approximations of the semantic interpretations humans might have given, on large scale and low cost. It is an empirical question, of course, how well the automated detectors approximate the human coders' tagging.

Sao Pedro, Baker, Gobert, Montalvo, and Nakama (2011), for example, describe their application of the technique to students' science investigations in a simulated microworld. Two of the authors tagged segments of students' action sequences as indicative of "Designed Controlled Experiments", "Tested Stated Hypothesis", "Used Data Table to Plan", and "Used Hypothesis List to Plan". A given clip might address zero, one, or more of these categories. They tagged a clip as "Designed Controlled Experiments," for instance, if it contained actions that suggested students were trying to isolate the effects of one variable. Clip tagging were the targets for prediction, and the potential predictors were counts and timing features of seventy-three automatically-detectable lower-level features, including variables changed when making hypotheses, hypotheses made, total trials run, incomplete trials run, complete trials run, pauses, data table displays, hypothesis list displays, and variable changes made when designing experiments. They used a decision-tree method similar to CART to define the detectors. The resulting classification functions could then in turn be used as triggers for feedback, and as values of observable variables for input to evidence accumulation processes.

Measurement Models

This chapter addresses using measurement models to synthesize nuggets of evidence (in the form of values of observable variables) across observations, in terms of student model variables (SMVs). As noted, this is not the only way that observable features of game situations and players' actions can be used in GBAs for either game or assessment purposes, such as tuning the situations or providing feedback. But it is a way to accumulate information across multiple sources of evidence, expressed as belief about characteristics of players, transitory or persistent. Further, it comes with tools to sort out evidence in complicated circumstances, quantify its properties, and assemble evidence-gathering and analysis components quite flexibly. We start by saying a bit more generally about these qualities of psychometric models, then look more closely at some models that seem to be particularly useful in GBA.

When probability-based psychometric models are used in evidence accumulation, reasoning is bidirectional. Their essential structure is modeling the probability distributions of observables conditional on SMVs—that is, reasoning from student characteristics to what we can observe. For example, we might posit that a student's probability of making an effective response in a family of similar situations is some unknown value p at this point in time, then observe particular instances. Or we might posit that a student at the 3rd level in a systems-thinking learning progression will give an explanation that is off-target (a level 0 response), one-variable-at-a-time (a level 1 response), allowing interactions (level 2), and incorporating feedback mechanisms (level 3) to be, respectively, *(p30, p31, p32, p33)*; we would expect the probabilities for more sophisticated responses to be higher than those of a student who is at level 1. Much of the history of psychometrics has been how to build such models and figure out how to estimate these conditional distributions.[7]

Once such a model is in place, the machinery of probability enables us to reason back the other way, from seeing a student's values of given observable variables, to updating our beliefs about the values of his SMVs (via Bayes Theorem; Mislevy & Gitomer, 1996). These updated beliefs can then be the basis of selecting activities, modifying features of game situations, providing feedback to students, and reporting summaries to teachers.

Both the nature and the grain-size (i.e., level of detail) of SMVs depends on their intended use—generally more detailed and qualitative for the tight feedback loops that monitor and guide play, broader for characteristics that change more slowly or pertain to more extensive definitions of proficiencies that provide summary results to players or teachers. There can be multiple feedback loops in operation in a GBA at a given time, which need to synthesize information at different levels

[7] An inherent challenge in using latent-variable models is that values of SMVs can never, by assumption, known with certainty, and inference must be through patterns among observed values in some collection of people.

of detail or at different time spans. For example, in a game with levels there can be psychometric models that operate within levels and others across levels; summary results within levels can provide information that updates the across-levels models.

The following sections say more about the forms and uses of measurement models that can be useful in GBAs. The next chapter, on psychometric properties, will say more about characterizing the weight and direction of evidence, checking whether the models are doing what we want them to, and making inferences when different people provide different kinds or amounts of evidence. In psychometric terms, these are questions of reliability, validity, and comparability.

Observed Score Models [aka Classical Test Theory]

Many games already use counts and timing data. We can apply familiar psychometric methods to examine the qualities of evidence that result. The premise of classical test theory (CTT) is that an observed score is the sum of a true score and random error. The count or proportion correct that is actually observed is viewed as arising from a conditional probability distribution given the true score. Several thought experiments provide a conceptual foundation for CTT (Lord & Novick, 1968), but this simple conception generates a surprisingly large array of useful tools for practical work to characterize evidence (Gulliksen, 1950/1987).

In practical terms, an average or sum across several observed scores is taken to approximate the true score, and more observed scores is better. "Reliability" quantifies the amount of true score relative to random error. The more closely multiple measures of a construct are related, the more reliable the average or sum is as a measure of it. In addition to sheer amount of evidence is its internal consistency: Multiple measures that tend to point in similar directions are more reliable than ones that give conflicting messages. When there are multiple opportunities to get information about a player's proficiency for a certain kind of task, even simple calculations of reliability tell us a lot about the quality of evidence.

In GBAs and particularly in the SimCity based challenges, a specific scenario is presented and test takers need to become familiar with the mechanics of the game. Furthermore, whenever a new challenge is presented, a new scenario needs to be introduced with, possibly, the requirement to learn additional game mechanics. This will all take up substantial assessment time and, therefore, the number of challenges that can be presented is limited. In addition, given the specificity of each challenge (which is critical in order to build an engaging, story-based experience), relatively few observed scores will be derived. Suppose the broad systems-thinking construct was only being measured by two observable variables in the power pollution challenge: the end state of the challenge (i.e., how much power and pollution is generated) and test takers ability to remove and replace the right objects and zones (i.e., coal power and heavy industries replaced with cleaner alternatives, much as wind/solar and commercial activities). The reliability will be much lower than, say, a 60-question multiple-choice assessment. However, accumulating evidence across challenges would provide a more reliable measure. In contrast, high reliability could be obtained for more specific constructs such as the ability to remove and replace objects, since a player must perform many remove/replace actions.

CTT works well when the multiple measures at issue are similar pieces of evidence about the same thing—in familiar assessments, for example, correctness across many similar test items; in GBAs, this would correspond to independent attempts at similar problems, as long as learning is negligible across those attempts. It doesn't work as well for situations that are more complicated in any of several ways: for example, where the evidence comes in different forms, has dependencies among some of its pieces, depends on different mixes of skills in different combinations, proficiencies are changing across the course of observation, or different players contribute different amounts or different types of evidence. Latent variable models were invented in psychometrics to deal with assessments with these features.

Latent Variable Models

In contrast to observed score or classical test theory models, a different class of measurement models has been developed based on the conceits that the variable denoting the construct of interest is not directly observable at all (i.e., latent) and that the relationship between observable indicators and the latent variable can be specified. The main ideas are these: We are interested in peoples' capabilities to act in some ways in a class of situations, and we can observe in each situation some salient features of their performance—these are the observable variables, or OVs, discussed in the previous chapter--denoted x_j for Task *j*, where x_j could be vector-valued.

We posit that students' performances, characterized by features x_j, arise from some underlying dimensions of knowledge, skill, familiarity, preferences, strategy availabilities, or whatever way we want to characterize them for the purposes at hand. These are called latent variables in the psychometric literature, and student model variables (SMVs), or sometimes competencies or proficiencies, in ECD terminology. We will denote them by θ, which also can be vector-valued. We model probability distributions for the observable variables for a Task j conditional on the latent variables, say $h_j(x_j|\theta)$. These are called conditional probability distributions, or sometimes more simply, links (Moustaki & Knott, 2000). Under appropriate conditions, we can estimate the conditional probability distributions, and given a person's observed responses x, make inferences about her θ based on the information they contain due to their probabilistic dependence on θ.

The forms of the θs, the x_{js}, and the links are determined by the nature and grain-size of the inferences about players that are needed, the forms of the observables, and the relationship between them. The relationships are determined partly by the design of the tasks, the conceptualization of how performance depends on the posited proficiencies, and to the extent it is available, actual data. We will be using an example from Jackson City for illustrations—specifically, the SystemModeling student model variable, and the MultivariateThinking, Accuracy, and JobsPollutionEndstate observable variables. First we review some additional key ideas in latent variable models and note how they are important in assessment and GBA.

The General Form of Latent Variable Models

As mentioned above, values of θ are by nature never observable. All we ever see is distributions of an observable variable, say $f(x_j)$, which in a given population is a convolution of the conditional probabilities given θ and the distribution of θ in that population, g(θ):

$$f(x_j)=\int g(\theta)\, h_j\,(x_j|\theta)d\theta$$

(Bartholomew & Knott, 1999). How can we estimate the link functions or make inferences about θs? The answer is by positing that the associations among the observable variables can be accounted for, for the most part, by the latent variables.

Expressed in words, if we proceed as if once we knew the values of a student's θ, the link functions would tell us what to expect up probabilistically for each of the items, and the actual pattern of observations among OVs beyond that doesn't depend on θ. The "as if" clause signals that we do not claim that the model is psychological truth, but rather a model to help us reason about students' capabilities, in ways we think will be useful, from the patterns in what we can see. It is this simplification that gives us enough information to both estimate the link functions and to make inferences about students' θs. What's more, comparing the patterns we observe among OVs with the patterns the model would predict enables us to modify the model if we need to, or the way we make observations (i.e., revise the game or the scoring rules), or even our underlying theories about what is happening in the game (Levy, 2006).

Expressed in statistical terms, the latent variable model posits conditional independence among the OVs across J tasks, $x\equiv(x_1,\ldots,x_j,\ldots,x_J)$; that is:

$$h(x|\theta)=\prod_j h_j\,(x_j|\theta)$$

Figure 18 depicts the conditional independence relationship graphically in a unidimensional model, that is, a latent variable model with just a single latent variable θ. The directed edges (arrows) from θ to each x indicate that each observable variable is modeled as depending on θ. The lack of edges among the xs indicates their distributions do not depend on other *xs* once θ is taken into account. The edges represent the link functions $h_j\,(x_j|\theta)$. Note that the direction of the edges points from the variable on the right side of the conditioning bar in the link function, θ, to the variable on the left side of the conditioning bar, x_j. The middle panel of the figure additionally shows that these conditional distributions can depend on possibly vector-valued parameters, denoted $ß_j$, for each Task *j*; we write

$h_j(x_j | \theta, ß_j)$. These parameters specify the relationship between θ and x_j, indicating properties such as the difficulty of tasks and how much evidence they provide about θ. (We will return to the right panel shortly.)

Figure 18:
Directed Graph Representations for an IRT Model

Task parameters that specify conditional distributions of xs given Θ are implicit in the left panel and explicit in the right panel.

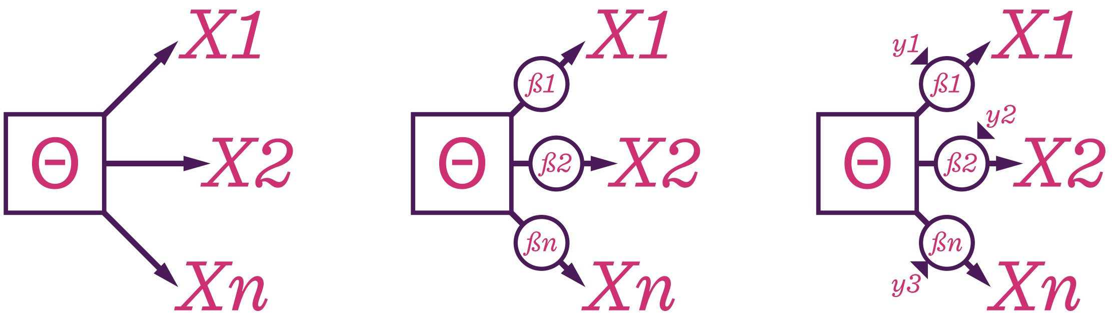
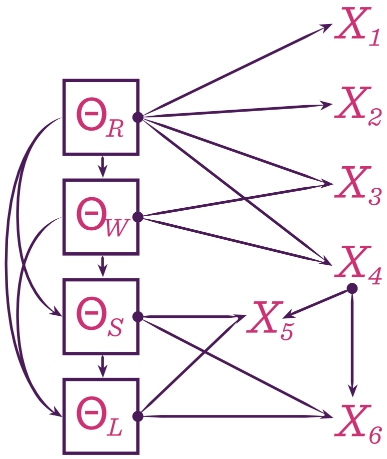


Figure 18
Based on Figure 1 from Almond, Russell G.; Mislevy, Robert J. (1998) ETS RR-98-04, TOEFL-TR-14, Graphical Models and Computerized Adaptive Testing. Reprinted with permission.

Figure 19 is the graph for a multivariate latent variable model, with four student-model variables. Three observables are conditionally independent, but x_4, x_5, and x_6 are inter-related aspects of a performance on an extended task and are more strongly related to one another than just their dependence on the θs would indicate (i.e., they are conditionally dependent). This model is similar to the unidimensional model in that the observable variables depend on the latent variables, and the observables from different tasks are conditionally independent given the latent variables. New features are relationships among the student model variables, and the fact that different observables depend on different patterns of student model variables.

Figure 19:
Directed Graph Representation of a Multivariate Model



Ultimately our interest lies in what can be said about θ given the observation of x in order to rank, order, predict, modify, report, provide feedback, or adjust a situation according to the construct of interest based on the observations. Reasoning back about θ given x can be expressed through Bayes Theorem as:

$$g(\theta|x)=(g(\theta)h(x|\theta))\,/\,f(x)$$

For example, we would like to know what the level of system thinking (θ) of a student is, given the observation that he completed the Jackson City challenge (x) within the given time limit. We posit an initial (prior) distribution for levels of systems thinking across all players g(θ). For Jackson City, our prior distribution will be mainly on levels 2 and 3, since it is students at these levels for which the game is intended. We observe the value of the observable variables x arising from a player's actions, and calculate revised beliefs about her likely status at the different levels of the Systems Thinking student model variable θ in light of her performance. This reasoning is the same regardless of the particular form of the model. A small numerical example will be given shortly for a Bayesian inference network, or Bayes net, psychometric model.

We will note two more general features of latent variable models, then discuss some particular models that can be useful in GBA that have these properties. The relevant features are the incorporation of situation features into the model and modularity of model construction.

Incorporating features of situations into the models

The parameters $ß_j$ in link functions that indicate how observable variables depend on student model variables can in turn depend on features of the task situation, say y_j; we then write $h_j\,(x_j\,|\theta, ß_j\,(y_j\,)\,)$. This relationship is depicted in the rightmost panel of Figure 18. These y_js are "data concerning the task" in the assessment argument (Figure 6) from an ECD perspective, and paradata from a data analysis perspective. Initial work exploiting these kinds of relationships in assessment appeared in the 1960s (e.g., Suppes and Morningstar, 1972), and formalized by Fischer (1973) in his linear logistic test model. By now numerous extensions of the ideas and rigorous statistical inferential procedures appear in publications such as Geerlings, Glas, and van der Linden (2011).

Incorporating features of situations into latent variable models can be important for modeling observable variables in predetermined work products. However, the move is critical for modeling observable variables from contingent work products: A vector of situation features is used both to (1) identify a situation where an instance of a pre-identified class of evidence-bearing situations occurs, and (2) indicate the probability model that applies, as to what the observables will be, which student model variables they depend on, and the nature and strengths of the relationship.[8]

[8] For a given vector of situational features, distinct, possibly overlapping, sets can be relevant for the purposes of identifying situations, indicating the form and variables in the necessary model fragment, and specifying the link functions.

In standard applications of latent variable models such as item response theory (IRT—more about this below), items are constructed individually, considerable data are collected for each of many items, and the $ß_j$ in their link functions are estimated uniquely and accurately for each item individually. This can be done in GBAs also for OVs that come from predetermined work products that all players, or at least many players, all take in the same form. Contingent work products, however, arise uniquely from players' actions. All instances of a given contingent work product with the same y values belong to the same equivalence class of tasks, although they may differ as to specifics that are not captured in y. Technically, in Bayesian terms we are considering them as exchangeable (Lindley & Novick, 1981): we might imagine they each had their own value of $ß$, even though we can't get repeated observations on each one to estimate them. We can however consider the probability distribution of a response x, conditional on θ, for a random member of the task class. If we are willing to assume that the instances of a contingent work product in a given data set are representative of the class, treating all the instances of OVs from the same contingent work product class as if they were the same task in an estimation program produces what is called an expected response function: A link function that approximates the expected value of response given θ as averaged over the class (Lewis, 2001).

There is a plus and a minus to being able to use expected response functions to model OVs from contingent work products. The plus is that we can in fact do it—this is remarkable in itself, because it says we can use psychometric machinery originally invented for individual test items with masses of data for each, but now in individual, unique situations that arise in situations determined at partly by players themselves. This is possible, it must be emphasized, only through design and theory. The theory indicates how actions in circumstances with what kinds of features ought to depend on students' capabilities. The design comes from arranging the simulation or GBA contexts so that instances of these kinds of situations can arise, and we have been able to characterize the particular features y that will alert us to when this happens. We see a convergence among measurement modeling advances, design strategies, theory about the domain and learning in the domain, and digital capabilities to produce then recognize these instances. The minus is that the information is attenuated from what it would have been had all players encountered the same exact situation, and we could have estimated its unique parameters. The loss of information depends on how much variation there is among the different members of the contingent-work-product equivalence class. If they do in fact all tap the same competencies in pretty much the same way and the variation can be modeled quite well by y values, then little information is lost. The more variance within an equivalence class, the less information each instance provides (Mislevy, Sheehan, & Wingersky, 1993).

The latent-variable framework and Bayesian paradigm enables us to take advantage jointly of information from theories about a learning domain, from our design strategies, and accumulating data from players. At the beginning, we posit models that reflect our initial beliefs about the targeted aspects of proficiency and the features of situations that will evoke them in various combinations. We build these hypotheses into the forms and the parameterizations of the models. By modeling conditional probabilities in terms of parameters, we can express our initial expectations as prior probability distributions for the ß parameters—ideally as intuitive functions of the situation (task) features y. As data arrive, Bayesian machinery allows us to get increasingly improved estimates of the ßs. Perhaps more importantly, we can examine where and how well the data fit the models, and where unexpected patterns arise. This information helps us fine-tune models to better manage evidence, or to modify game situations to provide better evidence.

Modular assembly of latent variable models

Latent variable models are particularly well-suited to assessments such as games and simulations that are interactive and allow students to work down different paths because of the conditional independence structure (Equation 2). The student model variables θ in a given student model are of interest for some period of time, whether a local challenge, a level, or the game as a whole. The link functions for observations are used to update beliefs about θ as new data x arrive in batches or in sequence via Bayes Theorem (Equation 3). The key point is that this process holds even when different players have different sets of observables, and even if the assessment situations have been presented based on a player's previous actions, such as presenting more challenging situations to players who are doing well and easier challenges to players who are struggling (Mislevy, in press).

In ECD terms, we see this happening in the four-process delivery processes. The presentation process in which the player interacts with the game produces work products – sometimes predetermined work products, sometimes contingent work products as they arise, sometimes from log files at checkpoints during play or at natural junctures such as submitting a solution to a challenge. Evidence-identification processes are applied to obtain the values of observable variables. In addition to their uses for immediate feedback or game adjustments, they are passed to the evidence accumulation process, where they are used to update beliefs about the student model variables. In an evidence-accumulation process based on a latent variable model and Bayesian updating, the process takes the form of docking appropriate link functions to the student model and carrying out Bayesian updating (Almond & Mislevy, 1999; Mislevy, Steinberg, Breyer, Johnson, & Almond, 2002). Figure 20 suggests this modular model-construction and updating machinery using the multivariate model depicted in Figure 19.

Again the latent-variable framework and Bayesian paradigm offer a particular advantage to complex assessments such as GBAs and simulations when we expect to improve the system over time: it is encapsulation of evidence management. As long as the student model variables remain the same, it is straightforward to incorporate additional forms of evidence, such as new observables discovered from mining accumulating log file data, or observables from new challenges, or additional observables from existing challenges.

Further, changes to evidence accumulation caused by game feature modifications or from improved evidence identification routines are isolated in the forms and parameters of link functions of observables from just the affected work products. The impact of ongoing system changes to evidence accumulation is thus easier to manage and keep coherent with the bidirectional reasoning in probability-based paradigm than one based on alternatives such as fuzzy logic, rule-based systems, neural nets, or confidence-factors (Spiegelhalter, Dawid, Lauritzen, & Cowell, 1993)—even though these approaches (even different ones for different work products and observable variables) can be quite satisfactory for the one directional reasoning needed in evidence identification.

Figure 20:
The Student Model and a Library of Links That Are Used to Update Beliefs About its Variable Whenever Evidence Arrives in Various Forms

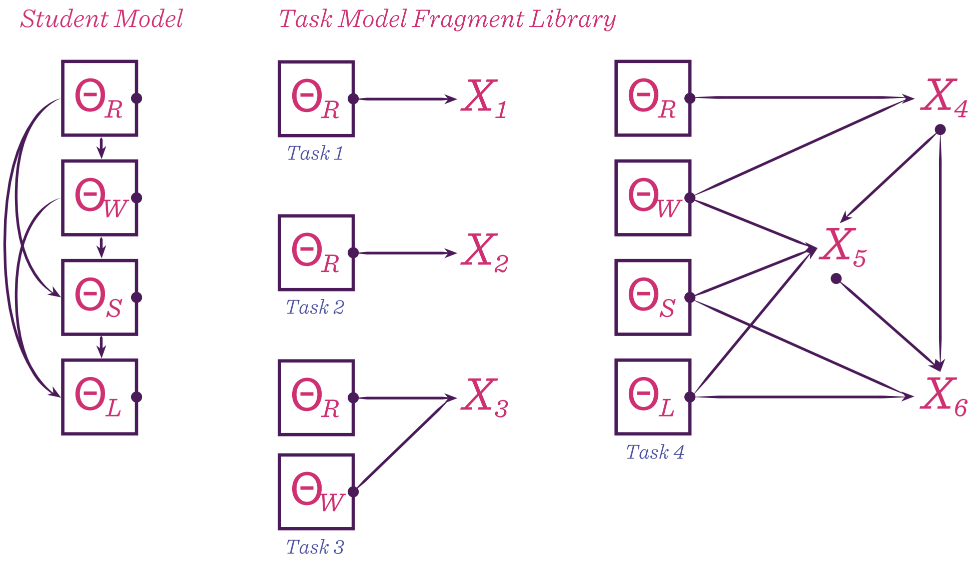


Figure 20
Based on Figure 7 from Almond, Russell G.; Mislevy, Robert J. (1998) ETS RR-98-04, TOEFL-TR-14, Graphical Models and Computerized Adaptive Testing. Reprinted with permission.

Some pertinent latent variable models

There are a number of psychometric models that have the properties discussed above. This section discusses three classes of such models. It should first be noted that modular model-building, estimation, and inferential methods have become the new paradigm in the world of statistics (particularly Bayesian statistics; see, for example, Clark, 2005, and Gelman, Carlin, Stern, & Rubin, 2004). Although it is useful to group models to discuss variations that can be exploited in GBA, there is no need to pick “a” model from among them to use. It is possible to mix and match these ideas for both kinds of latent variables and kinds of observable variables, just with appropriate choices of link functions (see De Boeck & Wilson, 2004; Rupp, 2002; and M. von Davier, 2005). Furthermore, we have noted that it is sometimes useful to have multiple versions of the four-process delivery system running in parallel, or to have them running at different levels of the hierarchical organization of GBA interactions. In both senses, the models can be of different types and different mixtures of model components.

Item response theory (IRT; Yen & Fitzpatrick, 2006) was originally developed for scoring students and modeling item-level performances on achievement tests. The form is that of Equation 2, with the xs being responses to dichotomous right-wrong test items, the θ a real number indicating a student's overall proficiency in a domain of items, the link functions being normal or logistic cumulative distributions, and the item parameters *ß* indicating properties such as difficulty. Through the modularity property discussed above, IRT made it possible to tailor tests to individuals, presenting items of appropriate difficulty to each examinee in light of her responses as the test progresses.

Extensions over the years support a wider variety of observable variables that might occur in games and assessments, such as counts, response times, ordered and unordered categorical variables, and sets of conditionally dependent responses (i.e., testlets; Wainer, Bradlow, & Wang, 2007). In Jackson City, if SystemsModeling were the only student model variable and all the observable variables were ordered category variables like MultivariateThinking, a partial-credit type IRT model could be employed to model performance. The resulting SMV would be continuous, and successively higher regions of it would correspond to the different levels distinguished in its conceptual definition.

Two extensions of IRT are particularly important to games, simulations, and complex performances more generally. The first is accommodating multivariate θs, to address multiple aspects of proficiency that are required in different mixes in different situations (as in Figure 19). These are called multidimensional IRT, or MIRT, models (Reckase, 2009). The second is formal incorporating item features y, to model item parameters *ß* (De Boeck & Wilson, 2004; Geerlings, Glas, & van der Linden, 2011). These have been called structured IRT models. An example of an IRT model with both of these characteristics is Adams, Wilson, and Wang's (1997) multidimensional random coefficients multinomial logit model. This move connects cognitive theory, task design (or discovery), and psychometric modeling, and supports modularity of model building and model use, in the ways described above. In particular, Wilson and his colleagues have been developing ways of relating regions of continuous IRT and MIRT variables with levels of stage-like learning for some time (e.g., Wilson, 1989), and more recently for learning progressions like the one for systems thinking in Table 2 (Wilson, 2009, 2012).

Diagnostic classification models are a particular class of multivariate latent variable models that often involve many categorical latent variables and many observed indicators that are indicators one or more of the latent variables (Rupp, Templin, & Henson, 2010; M. Von Davier, 2005). These models are particularly useful when there are several constructs at play that are entangled in tasks, and inferences are desired in terms of categories such as mastery, partial mastery, and non-mastery of those constructs. The complexity of the observed-to-latent variable relations is generally traded off against having simpler latent variables (e.g., mastery versus non-mastery, as opposed to ordering students along many scale score points). A key tool for these models is a Q-matrix. This matrix indicates which observable variables relate to which latent variables, usually in many-to-many relationships. This matrix is determined by task construction, with the elements of the Q-matrix being functions of paradata y about item features. There are both exploratory and confirmatory approaches to developing a Q-matrix, but some starting point and separation between latent variables is required in order to create a solvable problem.

Bayesian inference networks, or Bayes nets for short, are a broad class of models for interrelationships among categorical variables. They can be applied as psychometric models by operationally defining observable variables that depend on unobservable student model variables in the structure of Equation 1 (Almond & Mislevy, 1999; Mislevy & Gitomer, 1996; VanLehn, 2008). This makes Bayes nets a kind of latent class model, in terms of the history of psychometrics (Dayton, 1998)—in particular, when we parameterize conditional probabilities with parameters *ß* and model them in terms of ys, a structured latent class model with concomitant variables.

The general advantageous properties discussed above hold. Particular advantages of Bayes nets are great flexibility in the kinds of relationships that can be modeled and rapid updating of beliefs as evidence arrives, as might be required for making real time decisions in the presentation of game and simulation conditions. Koenig, Lee, Iseli, and Wainess (2010) and Shute (2011) illustrate the use of Bayes nets in game-based assessment, with ECD as the design framework. VanLehn (2008) provides a good overview for related uses in intelligent tutoring systems. The following numerical example provides some insight into how Bayes nets work in assessment. As one of the “interesting” issues for psychometrics in GBA, Chapter 11 will include discussion of dynamic Bayes nets to model change of students’ capabilities over time.

A Numerical Example

Figure 21 gives a numerical example of a part of a Bayes net for Jackson City. SystemsModeling is the latent student model variable (SMV) and MultivariateThinking and Accuracy are observable variables. Recall that SystemsModeling has five levels, labeled 1 through 5. Panel A of Figure 21 shows first the vector of prior probabilities we assign to a student being at these levels, before observing her performance. This is $g(\theta)$ in Equation 3. The values shown there represent beliefs that correspond to how we expect the game to be used: At this point, we anticipate most players would be at Levels 1, 2, or 3 (with respect to this context and content), and not at 4 or 5. In the matrix labeled MultivariateThinking, each row represents the conditional probabilities of observing performances coded 0, ..., 5 for a player at the level corresponding to that row. These are the values of $h_j(x_{\mathrm{j}}|\theta)$, where the j denotes the MultivariateThinking observable variable, and the entry in each cell corresponds to a possible value xj in that column given the person is at the θ for that row.

We see that the conditional probabilities for players at lower levels of SystemsModeling are modeled as being more likely to give responses at lower levels of the MultivariateThinking aspect of diagramming. As we look at increasingly higher levels of SystemsModeling (successive rows down the matrix), we see a shift toward higher conditional probabilities for producing higher values of MultivariateThinking. The matrix labeled Accuracy shows similar patterns for the conditional probability distributions (the rows) of values for the Accuracy observable variable, given values of the same SystemsModeling student model variable. We also note the spread of these conditional probabilities: Occasionally players at low levels give high responses and vice versa. The tightness of these conditional probability distributions will contribute to the strength of inference these observable variables afford about a student's SystemsModeling value.

So these matrices indicate an observer's (or the system's) knowledge before observing the student's performance: Prior beliefs about the student's level, based on background knowledge about the situation and the player, and conditional probability matrices that express performance expectations for players at each of the levels. We will say a more shortly about where these numbers come from.

Equation 3 implies that once a particular value of an observable variable is ascertained, we read down the column of the appropriate conditional probability matrix to see how likely that response was at the different possible levels of the student model variable. Their relative values tell us how to shift our beliefs from g(θ). These are numbers from the conditional probability distributions $h_j(x_j|\theta)$, but now x_j is fixed at the observed value and the column is a function of the unobservable SMV θ. In technical terms, the column corresponding to a particular value x_j is the likelihood function for θ induced by the observation of x_j.

Panel B of Figure 21 shows how Equation 3 is calculated when we observe values of 1 for both MultivariateThinking and Accuracy. We begin with the prior distribution g(θ). We multiply that vector, element by element across the rows for the possible values of θ, by the likelihood functions induced by the observed values for both observed variables. (The conditional independence form, Equation 2, says that updating beliefs takes the form of multiplying the likelihoods.) Multiplying across a given row gives adjusted beliefs, reflecting the strength of initial belief and the degree of revision from each of the two observations. The resulting column labeled Products reflect the relative strength of our beliefs about the player being at each possible value of θ after we see and evaluate her performance. They don't add up to one, as probabilities need to; their sum is the so-called marginalization constant in the denominator of 3 (which is what Equation 1 boils down to in this simple example). Dividing the products through by this number gives the posterior probabilities $g(\theta|x)$. From this performance, we see beliefs shifted somewhat down to lower range of SystemsModeling.

Panel C of Figure 21 shows analogous calculations for obtaining values of 3 and 2 respectively for MultivariateThinking and Accuracy. Compared to initial beliefs, the posterior probability is shifted toward higher levels. Level 3 seems most likely, but there is appreciable probability for Level 2, and nontrivial belief even about Levels 1 and 4. If we wanted to be more certain, we would obtain more evidence in the form of observing additional performances. The likelihoods induced by the values of the observables from that work would start with the $g(\theta|x)$ previously obtained, so it would then play the role of the prior for subsequent inference.

Figure 21:
A Numerical Example of Bayes Nets

Prior Probability and Conditional Probability Table

SMV Level	Prior Prob.	Multivariate Thinking					Accuracy		
		0	1	2	3	4	1	2	3
1	.25	.365	.259	.170	.118	.088	.454	.321	.225
2	.30	.220	.238	.221	.178	.143	.351	.342	.307
3	.30	.143	.178	.221	.238	.220	.307	.342	.351
4	.10	.088	.118	.170	.259	.365	.225	.321	.454
5	.05	.050	.113	.150	.290	.397	.200	.300	.500

Calculating Posterior Probability After Observing (1,1)

SMV Level	Prior Prob.		Multivariate Thinking						Accuracy				Product		Post Prob
			0	1	2	3	4		1	2	3				
1	.25		.365	.259	.170	.118	.088		.454	.321	.225		.02940		.394
2	.30		.220	.238	.221	.178	.143		.351	.342	.307		.02501		.336
3	.30	x	.143	.178	.221	.238	.220	x	.307	.342	.351	=	.01634	∝	.220
4	.10		.088	.118	.170	.259	.365		.225	.321	.454		.00266		.036
5	.05		.050	.113	.150	.290	.397		.200	.300	.500		.00113		.015

Calculating Posterior Probability After Observing (3,2)

SMV Level	Prior Prob.		Multivariate Thinking						Accuracy				Product		Post Prob
			0	1	2	3	4		1	2	3				
1	.25		.365	.259	.170	.118	.088		.454	.321	.225		.00947		.146
2	.30		.220	.238	.221	.178	.143		.351	.342	.307		.01826		.282
3	.30	x	.143	.178	.221	.238	.220	x	.307	.342	.351	=	.02442	∝	.277
4	.10		.088	.118	.170	.259	.365		.225	.321	.454		.00832		.128
5	.05		.050	.113	.150	.290	.397		.200	.300	.500		.00435		.067

These ideas extend to situations where multiple SMVs are required in various combinations for performance in a situation, different combinations are required for different aspects of performance, and there are entanglements among aspects of performance (Almond & Mislevy, 1999; Mislevy & Gitomer, 1996). When there are multiple SMVs, relationships among them such as prerequisition and dependencies among levels of related learning progressions can be modeled in terms of conditional probability relationships among the SMVs (West et al., 2012). Observable variables can be modeled as depending on combinations of SMVs such as conjunctions, where knowledge and skill are required jointly, or disjunctions among sets, where a problem can be solved with different strategies that use different knowledge and skill. Conditional probability matrices for an observable now have rows for each combination of SMV "parents," and the conditional probability distributions reflect these relationships. Entanglements among observable variables such as being multiple aspects of the same performance or having one stage in a challenge depend on what was done in a previous stage can be modeled in terms of conditional probabilities for combinations of their values (Almond, Mulder, Hemat, & Yan, 2006; Beland & Mislevy, 1996).

Where do the numbers in g(θ) and the conditional probabilities $h_j(x_j|\theta)$ come from? In the previous paragraphs the patterns described for these probability distributions were justified in terms of what we knew about the situation—expectations based on knowing the kinds of students who would be players, familiarity with aspects of the content from the in-class activities that surround the game, research on systems thinking, and the features of the situations that are designed into the game. Prior distributions can be based on just such information, and at the beginning, this is all one has. The Bayesian framework, however, allows for coherent updating of the conditional probabilities as data arrive (Mislevy, Almond, Yan, & Steinberg, 1999). The numbers in the example are initial expert-opinion refined by data from a small alpha test. Further, this framework enables an analyst to compare the patterns in the data with the patterns the model can express, so that the model or the data-gathering situations can be improved (Levy, 2006; Williamson, Mislevy, & Almond, 2000).

It is clear from the example that there can be many numbers in these conditional probabilities, and a challenge to estimate. Moreover, even when we don't know quite what they ought to be, there are qualitative patterns we expect to see based on our theories and our design efforts, such as the expectation of increasingly higher performance at higher levels of proficiency, or jumps in probability at a level where understanding of a certain concept is needed to crack a challenge. We can incorporate this information by modeling conditional probability matrices in terms of functional forms that express these qualitative patterns and have parameters that capture the particular ways they play out with the tasks and the players from whom we obtain data (Almond, DiBello, Jenkins, et al., 2001; Almond, DiBello, Moulder, & Zapata-Rivera, 2007). This parameterization has the additional advantage of improving the stability of estimation.

This simple example illustrated a number of key ideas: Modeling salient aspects of students' proficiencies in terms of student-model variables. Modeling salient aspects of performance in terms of observable variables. Modeling distributions of observable variables in terms of conditional probabilities, given SMVs. Building and parameterizing the models in terms of theory, experience, and designed-in expectations. Using a Bayesian modeling framework so we can make coherent inferences about players, update the models as data become available, and assemble model fragments to suit evolving game situations. These same ideas obtain in exactly the same way conceptually with MIRT models and diagnostic classification models, even though the forms of the models and the details of calculation differ accordingly.

Re-Usability and Latent Variable Models

A pervasive lesson from experience with complex technology based assessments is that it is a bad idea to implement complicated task situations and capture rich data, and hope that someone done the line will be able to figure out "how to score it." Designing from the beginning around assessment arguments, even if roughly at first, may seem difficult but is more apt to succeed (Bennett & Bejar, 1998). A later section will have more to say specifically about this in a rapid iterative design process for game-based assessments. Here we want to call attention to the value of re-usable elements that include psychometric model fragments.

Because figuring out how to craft complex situations, capture relevant evidence, and make sense of it is generally hard to do, once we figure out how to do it we should capture the solution in appropriate representations to adapt and re-use in future situations. In assessments, this means structures such as task models and design patterns (e.g., Luecht, 2003, 2009). In a design framework for problem-solving in dental hygiene simulations, for example, Mislevy, Steinberg, et al. (2002) proposed scenario segments built around recurring situations such as conducting a patient history and choosing language appropriate to a colleague. These schemas indicated key task features to include and to vary, targeted student model variables, abstracted characterizations of observable variables, and Bayes net fragments that providing a skeleton for the link functions relating them.

In game based assessment, similar components, now connecting where possible with game mechanics, can be recognized in early work and similarly abstracted into re-usable structures. Game designers use this strategy already of course for game design. The object for GBA design is to have re-usable elements that have locally rectified some set of assessment design and game design considerations, to exploit in games that could look rather different on the surface.

“Interesting Factors” for Psychometrics in GBA

This section discusses a number of factors that arise in psychometrics, some only occasionally, others usually tacitly, that designers of GBAs will address regularly and explicitly. On the surface, psychometrics is about measuring latent variables, and measurement concepts and models are indeed central to their use. For GBAs (and other less familiar forms of assessment), it helps to view them more broadly as information-managing tools. [9] From this perspective we can see how design choices about psychometrics interact with design choices about learning and game play. We have already addressed some issues of this nature with regard to making sense of the rich and complex data that GBAs can provide. We now apply the same perspective to the nature and use of latent variable models in GBAs.

The first three factors discussed below concern the meaning of student model variables in psychometric models. While SMVs have labels that suggest a meaning and formal meanings in the model space. However, their effective meaning arises from their role in the assessment argument, which is intertwined with contextual grounding, design choices, and intended uses of information. The next three factors concern features of GBAs that are common in games but, to varying degrees, less so in familiar assessments. These are adaptivity, changing values of student competences over time among observations across time points, and collaboration among players.

What Else You Know Influences What Needs to Be in the Student Model

From the situative/sociocognitive psychological perspective, responding to even the simplest multiple choice item requires assembling myriad linguistic, cultural, and substantive patterns. Modeling responses with simple models only works (when it does) because of purposeful design choices for tasks and constraining (if implicit) determinations about the occasions of use and backgrounds of the people who are to be assessed. There are many potential meanings for the tasks and many ways of failing to interact with them in intended ways—including getting them wrong in ways that don’t fit the argument scheme for “what the test is supposed to measure.”

In order to “work” in the usual and generally assumed ways, then, many factors beyond the form of the assessment per se must be in place. These factors include the language and the mores of the assessment situation, the language structures and the genre of the test, the kinds of behaviors that are anticipated and how they will be evaluated, and common representational forms and common experiences – so that for the most part, the ways examinees differ is are mainly in line with the capabilities that users think the assessment “is supposed to measure.” All of this is usually implicit.

[9] Which is just as much the case in familiar assessments. There however they are sufficiently embedded in familiar situations and use cases that we can often use them fairly sensibly just by following standards of good practice and applying “common sense”—tacit knowledge built up over decades of what seems to work and what doesn’t in recurring situations in customary systems, accompanied by the formal machinery.

Tasks vary systematically across the space of the targeted capabilities (to achieve construct representation, in psychometric terms; Messick, 1989), and by design and presumption, differ minimally with regard to other presumed knowledge and skills of examinees (potential construct irrelevant sources of variance). For example, the test developer will try to keep the vocabulary and syntactic demands of chemistry tests much lower than demands for chemistry concepts, except for elements of language that are integral to the targeted chemistry concepts.

The more complicated task situations are, the more open-ended challenges are, and the more heterogeneous examinees' backgrounds and understanding are with respect to demand of the tasks other than the targeted ones, the wider the variety of interpretations can be for student's performances. This is the case of game-based assessments (and of simulation-based tests, performance assessments, and "authentic" tasks more generally).

If a game is about linear functions for modeling in an investigation, both linear-model proficiency and inquiry skills are involved—but if we already know a player is sufficiently familiar with linear models, we may only need to model the inquiry skills of interest. The situation is reversed if we know the player has considerable experience working investigations through the inquiry cycle schema, but what is new to her is doing so with linear models. If we know neither, our model might need to include SMVs from each of these aspects of the capabilities involved, to support inferences from noisy data about both kinds of proficiency. This use-case will be at once more complicated and less informative about either kind of proficiency.

Jackson City challenges are meant to formatively assess and develop systems thinking capabilities, doing so with a particular systems (pollution and jobs, and eleven other related factors), with particular representations and levels of English, in a particular SimCity-based environment, with expectations about what a solution might look like. There are many ways to go astray.

The meaning of the SystemModeling SMV in Jackson City, for example, is couched in terms of qualities of thinking while working with an unspecified interactive system. But systems thinking won't be evidenced at all if a student has trouble understanding the SimCity-style interface, or doesn't know that coal plants produce more pollution than solar plants, can't toggle between different views to get feedback on how her choices are affecting the city, doesn't know what "simoleons" are when she gets the message "You don't have enough simoleons to build a solar power plant."

For this reason, Jackson City provides in-game, little-g, help functions and feedback when problems are detected; but more importantly, the recommended use embeds game play within a larger big-G context. Students' game play is interspersed with ongoing teacher-guided activities that further support students in gaining background knowledge and skills required for success in the game. These more typical classroom activities are designed to introduce students to systems, systems thinking, and the accompanying vocabulary through use of real world contexts and introduction to causal loop diagrams. Student discussion, writing, and readings provide opportunities to learn how to use causal loop diagrams to identify the components of systems and reason about them. In the course of that work, students work in small groups to respond to real world policy scenarios that hinge on the same types of conflicts they must grapple with in the little g-game. In this way the time spent in teacher-guided activities gives students the tools and background knowledge they need to succeed in their little g-game play.

This contextualization through in-game supports and teacher guided classroom instruction supports interpretations of actions in the game space by developing the background knowledge and skills necessary for students to engage with the GBA tasks as they were designed. It increases the likelihood that the game will elicit evidence about students' systems thinking as opposed to effects from other nuisance variables such as how to work the mouse, which power plants generate the most amount of air pollution and why power plants impact the market for jobs, and so on. In this sense, the in-game feedback and the instruction surrounding the game are critical the inferences we want to make – namely, to be able to interpret probabilities over the SystemModeling variable as telling us something about a player's thinking about the system in the game.

The Situated Meaning of Student-Model Variables in GBAs

Another factor is related to the preceding discussion, but merits attention because of its close connection with the measurement concepts of generalizability and validity (which themselves will be discussed further in the next chapter). It is the fact that the meanings of student-model variables in any assessment application are grounded in the particulars of the observational settings and the persons whose performances are used to fit them model. Interpretations of scores in the form of summary statistics of student-model variables have this sense of meaning by construction. Whether they have additional senses—whether they ground inferences about other situations and/or other people—is an empirical question.

A situative, sociocognitive perspective on learning would urge caution, and would strongly advise against extrapolations based simply on a label attached to the SMV. The idea is that learning occurs in terms of resources developed in specific situations and is initially tied tightly to those situations (Greeno, 1998). Whether underlying concepts or capabilities would be activated in other situations depends on features of the new situations and whether the initial learning happened in ways that make that activation more likely—quite apart from parallels that might be readily apparent to an expert.

Empirical results from the performance assessment movement of the 1980s are sobering. Shavelson and his colleagues reported substantial variation in students' performance for closely parallel investigations on paper towel strength and sow bugs' preferences for light and moisture

(Shavelson, Gao, & Baxter, 1993), and even for simulation-based and hands-on forms of the very same investigation (Baxter & Shavelson, 1994). Ruiz-Primo and Shavelson (1996) concluded, “Whatever performance assessments are measuring about science understanding is highly sensitive not only to the task and occasion sampled, but also to the method used to assess performance.”

This is currently a pressing problem in science assessment, as initiatives such as the Next Generation Science Standards (NGSS; National Science Teachers Association, 2012) when assessment specifications call for the integration of disciplinary concepts and scientific practices. We can describe practices of developing and using models that are carried out with a great variety of particular models across a great range of situations, but it is arguable that there exists a unitary “model-based reasoning skill” in students’ heads which is applied conjunctively with particular models as they move from one context to the next. We can develop guidelines and design patterns to help test developers create assessments of model-based reasoning across various particular models and contexts (Mislevy, Riconscente, & Rutstein, 2009), but we can do this without having to believe that such a thing as “model-based reasoning ability” exists in people without regard to models and contexts.

The posterior distributions on the SystemsModeling SVM in Jackson City based on a players’ performance in this context are interpretive summaries of these particular performances, in this context, with the systems in play in the game. Whether (and if so, when and how) they hold meaning more broadly is an empirical question. One dimension for improving a GBA is designing the little-g game, the experiences, and the surrounding big-G activities so as to build resources that are apt to be activated in other situations (Hammer, Elby, Scherr, & Redish, 2005). It is probably expecting too much to think that one GBA can build up systems-thinking concepts in ways that are readily activated in a wide variety of applicable circumstances. On the other hand, enabling students to experience a variety of situations with different systems—all thought about, talked about, and investigated using the same concepts and representations—just might.

Reporting to Users

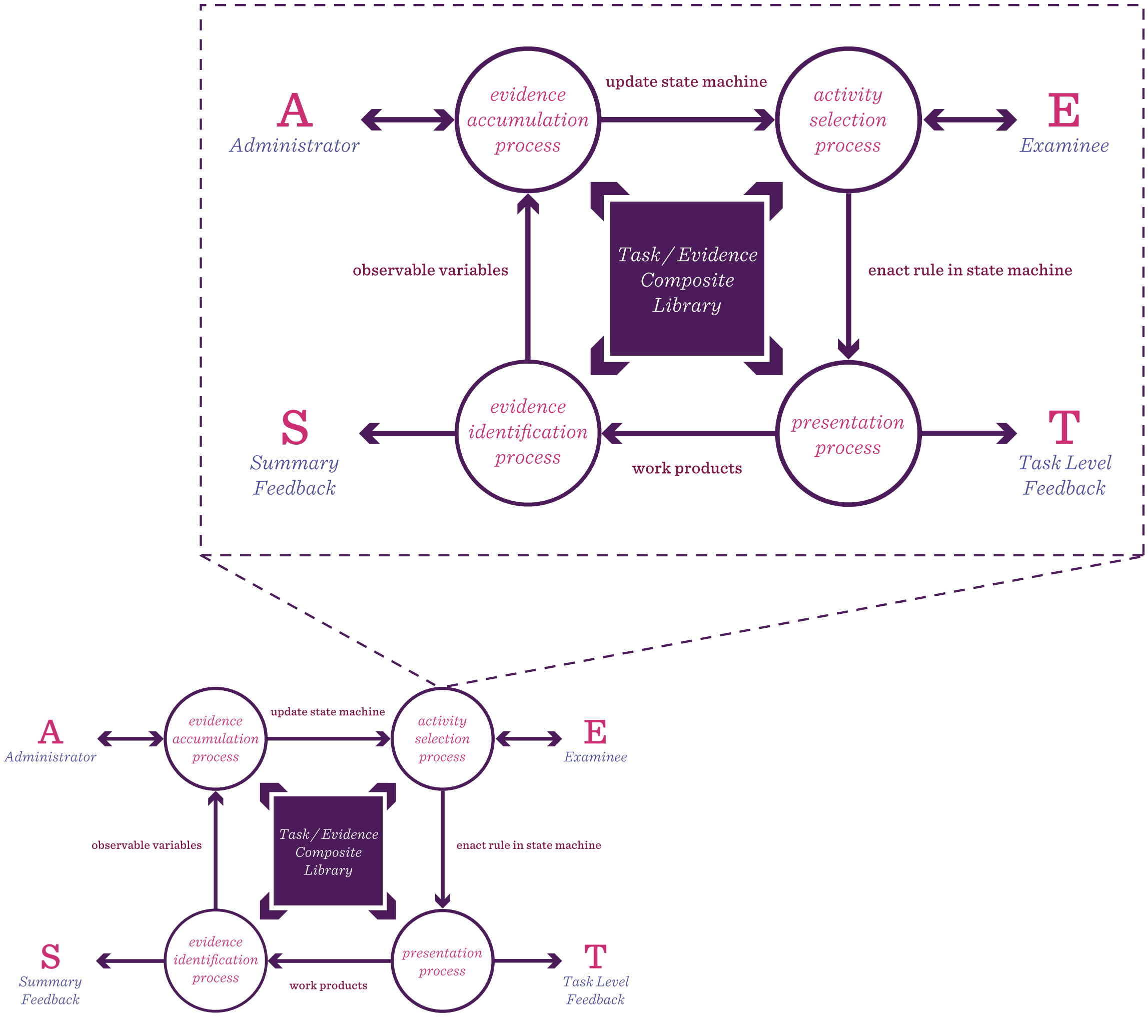
An important consideration for the success of familiar assessments is the user’s needs: Who needs what information, when, for what purpose(s), in what form? The design of the assessment shapes the assessment argument, then implements the machinery of the assessment, to support the use case that is at issue. Diagnostic tests usually need finer-grained student models, to give more focused feedback to either teachers or students themselves. End-of-course tests use fewer student-model variables, sometimes even just one to capture an overall proficiency, in order to gauge proficiency in a broad sample of tasks across capabilities the students have been studying throughout the course. Educational surveys such as the National Assessment of Educational Progress provide information to policy-makers and the public in terms of a relatively small number of curricular areas, based on large enough samples of students to give a good picture of the student population—but they do not collect enough data from sampled students to assess their capabilities or guide their learning individually.

Analogous considerations hold in GBA, with additional layers and complexities: There can be multiple users, interested in different questions, who need information at different grain-sizes and different time scales. The four-process delivery system (Figure 9) helps organize thinking. We will look at some possible users and needs in this section, and in the next section say more about the implications for Evidence Accumulation (i.e., measurement model) processes.

There are several points in the interactions among GBA processes that some agent needs information involving psychometrics. Moreover, as mentioned previously and will discussed further in the next section, there can be hierarchies of processing that are usefully thought of as nested four-process cycles, as depicted in Figure 22 (and effected by finite state machines). The kinds of activity and communication described below can take place at the same time at, for example, for the game as a whole, for levels or challenges within the game, and for more focused activities within levels or challenges. We will say a bit more shortly about implications of such hierarchical structures for psychometric models.

The presentation process controls interactions between the system and the player. In a computer-delivered multiple-choice test, for example, the presentation process renders and presents the information that appears on the screen as a test item, and recognizes and encodes an examinee's response—the work product, in this simple example. More steps of interaction are needed for an item that requires dragging-and-dropping icons to create a system diagram in Jackson City, but the idea is the same; the xml description of icons, locations, and links that exists when the students hits "submit" is the work product here.

Figure 22:
Nested Delivery Systems


A
Administrator
evidence accumulation process
update state machine
activity selection process
E
Examinee
observable variables
Task / Evidence Composite Library
enact rule in state machine
S
Summary Feedback
evidence identification process
work products
presentation process
T
Task Level Feedback
A
Administrator
evidence accumulation process
update state machine
activity selection process
E
Examinee
observable variables
Task / Evidence Composite Library
enact rule in state machine
S
Summary Feedback
evidence identification process
work products
presentation process
T
Task Level Feedback

Evaluations of work products by evidence identification processes, expressed as values of observable variables, can be considered psychometric work, and can be used for different purposes for informing different users. As depicted in the figure, the current evaluations can be used to trigger task-level feedback to the player, in the form of hints, encouragements, or explanations. Task-level feedback can also be passed to the activity selection process, to direct the presentation process to modify the game environment (e.g., decrease the difficulty for a player who is struggling). Values of observable variables can be included along with work products and traces of student actions in log files, for further analysis by designers and researchers.

The values of observable variables can also be passed to evidence accumulation processes. Here, as discussed above, information about performance expressed as observable variables is viewed as evidence about players' capabilities at some time scale, and synthesized as sums or counts in observed-score models or as posterior distributions over student model variables in measurement models. The summary feedback shown in the figure coming out to the left of the Evidence

Accumulation process can reported out to the player in a dashboard, either continuously, as a status report at the end of a challenge, or at the end of the full game. Summaries of SMV information across students can also be reported to teachers to monitor students' progress.

Information from the measurement model, whether full posteriors or simply as most likely values (Bayes mode estimates), can also be communicated to the Activity Selection process to trigger messages to the Presentation Process to modify the game environment—now based on evidence synthesized across multiple actions, rather than based on just the more immediate evaluations of particular actions coming directly out of Evidence Identification.

When there are hierarchies of delivery-process interactions, say at the levels of the game, challenge, and activities-within-challenge, the form of the models and the nature of the student-model variables can differ at different levels. The SMVs at inner levels might be of use only during that particular phase of play, being defined and monitored to understand a player's capabilities in order to provide feedback and adjust game features at just that level. At the end of that game segment, their values and the machinery for calculating them might have no further use, since simply noting completion may suffice, or their final values can be used to update coarser-grained SMVs at a higher level in the hierarchy.

It should also be noted that even at a given level of psychometric modeling, we do not need to pick a single model form or to model all incoming evidence in an all-encompassing model. It is possible to have multiple Evidence Accumulation processes running simultaneously for different purposes. For example, one can have models for observed-score mores for counts of certain events at the same time as a Bayes net model for evaluating work in terms of a learning progression, and at the same time having multiple latent class models running to accumulate evidence, if it occurs, for patterns among selected observable variable values that signal certain misconceptions or problems (e.g., a detector for lack of engagement; Baker, D'Mello, Ma.Mercedes, & Graesser, 2010). It is an advantage of partitioning processes and data objects in a GBA to be able to add and modify Evidence Identification and Evidence Accumulation processes in response to improvements in game design and learning from on-going data mining.

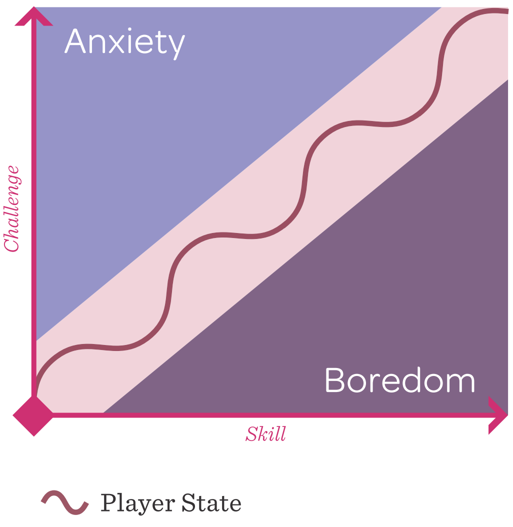
Adaptivity

The essential idea of adaptive testing can be traced back to Alfred Binet a century ago. It became more broadly adopted when computer based testing and item response theory became more commonly available. Adaptive testing is an approach to selecting test items, sections, or, more generally, materials based on test takers' ability gleaned from prior performance (Wainer et al., 2000). Prior performance can be based on previous tests (e.g., a screener test of some sort), information from external sources (e.g., grade level, or other test taker characteristics), or from earlier parts of the test. Adaptive schemes can be item based, where a test takers' level of performance is assessed after every item and the next item is selected based on that assessment, or section based, where performance on prior sections informs which section is administered next. The motivation for this kind of task selection centers around the statistical notion that the most accurate information about a test taker is obtained when the level of difficulty is close to the test taker's level of performance. However, there is also an important experiential aspect: test takers tend to perform best when items are just a bit challenging, but not too challenging. Items that are too hard demoralize the test taker, while items that are too easy bore her. Adaptive procedures for multivariate models (Segall, 2010) make it possible to select (or construct, or modify) tasks for a test taker that become more challenging in one aspect but easier in another, if this is what performance thus far suggests will keep her at the cusp of her capabilities.

Many games use an analogous strategy. The player is viewed as a learner who is continuously trying to level up, mastering skills incrementally. The graph in Figure 23 shows a timeline on the horizontal axis and skill level required on the vertical axis. As the player traverses the experience (solid diagonal serpentine line), she alternates between needing a skill level that is slightly above where she is at that moment (which will invite some anxiety) and mastering that skill while the game is about to add new challenges (which will invite some boredom). As the game subsequently asks a higher skill level from the player (e.g., a bigger monster, a more complex pollution problem, depicted by the dashed line), the player will alternately move between a state of learning and state of mastery. Being consistently in a state of anxiety or boredom turns a player away from the game, while making incremental learning accomplishments will motivate her. A good game adapts constantly to the skill level of the player without making it too hard or too easy for a sustained amount of time. This is similar to adaptive testing from an experiential perspective, and compatible as well with Vygotsky's (1978) famous notion of the zone of proximal development.

Results from three distinct lines of research thus converge: Experiences around the leading edge of one's capabilities optimize learning, assessment, and engagement. This is an aspect of GBAs where design principles from learning, psychometrics, and instruction work together.

Figure 23:
Chart for Optimal Experience



Changing Values of SMVs

Most educational assessments presume that the capabilities being assessed remain constant over the course of observation, and use measurement models that embody this assumption. An immediate implication of Figure 23, however, is that we can expect at least some aspects of players' capabilities to increase as they play a game. This means we need models that accommodate the possibility that the values of unobservable student model variables will be changing over time. Psychometric work concerning the dynamic testing paradigm (e.g., Embretson, 1990) and the learning models in tutoring systems (e.g., Corbett & Anderson, 1995) are areas we can be drawn on for GBAs designed to develop targeted proficiencies. Four basic approaches are listed below.

Recency-weighting of evidence. A first strategy is using psychometric models that do not accommodate change, but fading the influence of data as it recedes into the past; that is, recency-weighting evidence. A substantial advantage to this approach is that simpler models can be used. A disadvantage is that as the value of the latent variable changes, a current estimate from recency-weighted data lags behind the true current value. The trade-off in how aggressively to fade past data is that a shorter window makes the estimate more current, but a longer window provides more evidence and dampens noise. Two basic methods for down-weighting past data are these:

- Re-estimate a statistic whenever it is required using weights w(t) for each data point, with $w(t^*)=1$ for the current time point and $w(t)<1$ for $t< t^*$; for example, $w(t) = c^{(t^*- t)}$ for some fading constant $c < 1$. In this method, data x_t-1, x_t-2, ... must be retained for as long as the look-back window requires.

- Use the Bayesian updating scheme (Equation 3), but down-weight the prior distribution each time. Rather than , use:

$$p^{*}\left(\theta \middle| x_{t'}, x_{t'-1}, \ldots, x_1\right) \propto p\left(x_{t'} \middle| \theta\right) p^{*}\left(\theta \middle| x_{t'-1}, \ldots, x_1\right)$$

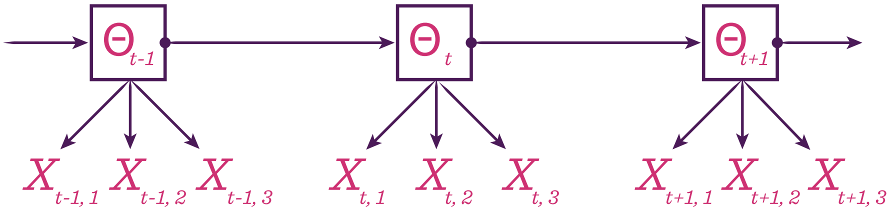
  where $p^*(\theta| x_{t'-1}, ..., x_1)$ is a weakened variant of $p(\theta| x_{t'-1}, ..., x_1)$. For example, in a problem where each $p(\theta| x_{t'-1}, ..., x_1)$ is a normal distribution $N(\mu,\sigma^2)$, use $N(\mu, c\,\sigma^2)$ with fading constant $c>1$. This method is well suited to IRT and MIRT models.

Bayesian model-tracing. A second strategy is the model tracing approach described in Corbett and Anderson (1995) and subsequent refinements and extensions (e.g., Baker, Corbett, & Aleven, 2008), used in a number of cognitive tutoring systems. This approach evolved from classical work in mathematical psychology, such as power-law learning curves and reinforcement models. When applied in its most basic form, it concerns a learner's repeated attempts to essentially equivalent dichotomously-scored problems. There is an unobservable probability—a latent variable—that she has "mastered" the skill in question, but a guessing probability of getting it right even if she has not mastered the skill and a "slip" probability of getting it wrong if she has. At the beginning of observation the analyst holds an initial probability that the student has mastered a skill, and on each attempt, a non-master has a probability T of moving to the mastery state.

This strategy has proved successful in a variety of cognitive tutors. In its basic form, it is applicable to GBA situations with focused, exchangeable, tasks. Further extensions would required for broader use in GBAs such as Jackson City, where there might not be crisply-defined tasks, task situations may differ in their difficulties, and different combinations of knowledge and skill may be required. Theory-driven situation design and resulting paradata are key to such an extension (Embretson, 1990). Whether for predefined work products or contingent ones, theory about the capabilities required in a given situation can be modeled in terms of features of those situations, as described above in connection with structured MIRT, diagnostic classification models, and Bayes nets. That is, theory and task design (or in the case of contingent work products, discernment) indicate which SMVs are involved, how they combine, and how much demand there is for each and for their combinations. When SMVs and OVs are both dichotomous, the resulting model is a dynamic version of diagnostic classification modeling.

Dynamic Bayes nets. Dynamic Bayes nets with latent student-model variables are Hidden Markov Models (HMMs). Figure 24 is a dynamic version of the Bayes net shown earlier as Figure 18. As before, the observable variables at time t, x_{tk}, depend on the unobservable value of the SMV at time t, θt. Additionally, the value of the SMV can change from one point to another, and is dependent on the previous value through the transition probabilities indicated by the edges from each time point to the next. It is further possible to condition the transition matrix on intervening experience, such as whether a player doing poorly at a given level chooses to take advantage of help that the system suggests or declines it. Examples of dynamic Bayes nets for modeling activity and learning in interactive environments appear in Iseli, Koenig, Lee, and Wainess (2010), Rowe and Lester (2010) and Ting, Phon-Amnuaisuk, and Chong (2008). Levy (in press) goes into further detail for modeling and estimation methods, for learning in an algebra GBA for middle school students called Save Patch (Chung, et al., 2010).

Figure 24:
A Dynamic Bayes Net



Periodically updating higher-level models. A fourth strategy, one we may pursue in Jackson City, is appealing when student modeling takes places in hierarchies. During a certain segment of play, a static student model and Evidence Accumulation process synthesize capabilities within that segment and adapt play or provide feedback. When the segment is completed, the fact of its completion, the degree of success, or the number of attempts is used to update beliefs about coarser SMVs in a higher-level model. Kimball's (1982) calculus tutor was an early application of this approach.

Multiple Attempts

A student can play a SimCityEDU challenge like Jackson City multiple times. Each time she experiments with tools and strategies, and gets feedback from the game explicitly as messages ("The city doesn't have enough power!") and implicitly through what happens as a result of her actions (jobs meter, pollution map, sims' comments). Each time, she may understand the system a little better, and have a better idea of how her actions reverberate through the system. An experience like this is less like a standard assessment than a "dynamic assessment" (Campione & Brown, 1987; Poehner, 2008), where rather than seeing how well a student can do in an unsupported attempt, we see how much and what kind of support it takes to for her to reach a given level of success. How should we handle multiple attempts in psychometric modeling?

As with many design decisions in GBA, the answer depends on what we want to do with the information. More specifically, in what way does the information in the number and character of attempts constitute evidence for some inference, for some user? We will consider a number of ways, some of which are implemented in SimCityEDU.

Simply treating the multiple attempts as providing several conditionally independent responses makes sense only when we expect the underlying proficiencies to be relatively constant. This raises a question, though, about just what we want to consider "the underlying proficiency" to be. There are different conceptualizations, connected with different target inferences. In particular, we can consider inferences local to SimCityEDU and inferences marginal with respect to systems modeling more generally. [10]

Considering inference local to SimCityEDU means examining systems modeling capabilities with respect to the set of systems that are at issue in the SimCityEDU challenges. It is an empirical question as to whether repeated attempts at, for example, Jackson City, result in increased effectiveness in interacting with the system in ways that meet the challenge of reducing pollution while maintaining jobs. It is an empirical question as to whether players' system diagram before and after successive attempts provide improved representations of the relationships among elements of the system.

We have seen enough data already to answer these empirical questions: Yes, almost all players do get better on repeated attempts, not only in solving the challenge but in modeling the system. We should not use a psychometric approach that assumes no increase in proficiency, at least locally. The modeling approach that is implemented as this is written is to enter a student's best attempt into the Bayes net. [11] The result is a characterization of the level of systems thinking that is represented their most effective performance.

[10] Inferences about other targets could be considered as well. We could consider noncognitive aspects of play such as persistence or engagement for example. Measurement of variables such as persistence not only benefits from multiple time points, but would seem to require them. For example, we might fit a survival curve that describes students' likelihood of replaying or not returning to the game as a function of previous success and number of previous plays. In other words, how successful are we in holding the player's attention within and across attempts?

[11] This is not always the last attempt. We have observed students work until they meet a challenge successfully, then play the challenge again to explore other aspects of the scenario. For example, some students figure out how to reduce pollution and maintain jobs—and then try to maximize the Sims' happiness under these conditions!

One of the psychometric approaches for changing SMVs discussed above would use more of the data without assuming constant proficiency, for locally or conditionally interpreted SMVs. Another alternative that could be employed with particular observable variables is to incorporate multiple attempts into a cross-attempt evidence identification process—that is, using a vector of scores on each particular observable across multiple attempts as an intermediate work product, and producing a graded response observable variable. [12] As a simple example, a dichotomous success/failure variable observed across multiple attempts could be converted to a multiple-attempt variable with values like "successful on the first try," "successful on try 2 or 3," "successful after more than 3 tries," or "unsuccessful." (Further distinction might be useful among "unsuccessful" as to number of attempts as well, to capture evidence about giving up quickly.)

Considering inference to systems modeling proficiency more generally, the question is whether repeated attempts and usual increases in local proficiency translate to improved understanding as it might apply more broadly. We are not modeling this in SimCityEDU, but it is worth considering how one might go about doing so. For when local learning is present, we want to understand if that learning transfers from the specific game and environment to more general capabilities, and where it does, describing its nature.

The key to investigating the presence and the extent of transfer would be having tasks that obtain evidence about system modeling outside the SimCityEDU environment. This might be done with a number of tasks concerning other systems a student could understand quickly and interact with, model, and solve problems in a more limited and time-constrained way than SimCityEDU does. We would want enough of them to calibrate as a broadly-conceived systems modeling SMV defined, say θ_{SMG} – proficiency in systems modeling generally, ideally defined through the same generically defined learning progression Table 2. For a sample of students, we would collect data for θ_{SMG} after the SimCityEDU experience. From the SimCityEDU experience, we would obtain both a final θ_{SML} – proficiency in systems modeling locally – and the data across multiple attempts such as number of attempts per challenge and mean θ_{SML} at each attempt. We could then carry out the following investigations:

1. Calibrate final θ_{SML} into the θ_{SMG} scale, using its relationship to Form B posttest results. This tells us the relationship of final performance levels for θ_{SML} to systems modeling proficiency more generally.

2. Calibrate the more detailed multiple-attempt counts and proficiencies from SimCityEDU into the θ_{SMG} scale. This tells us whether, and to what extent, evidence about learning while playing SimCityEDU provides information about systems modeling proficiency more generally. For example, the same final θ_{SML} level may be indicative of higher θ_{SMG} if it is achieved with fewer attempts rather than more attempts.

[12] An identification process can take as input both work products as provided by a presentation process and observed variables provided by previous evidence identification processes. Natural-language essay rating, for example, can use three or more passes by different EI processes, which operate at different grain-sizes and can take output of previous processes as their inputs.

The preceding investigations allow us to correlate evidence from SimCityEDU onto θ_{SMG}. It would be of particular interest to see the extent to which θ_{SML} might be an over-estimate of θ_{SMG}.

Suppose we have enough general tasks for two test forms, A and B, on the same scale, and gather pretest and posttest data from SimCityEDU players and one or more comparison groups of students at the pretest and posttest occasions. Between test occasions the other groups may do something unrelated to systems thinking (control group) or engage in a different activity related to systems thinking (treatment comparison groups). Another investigation bearing on a another inference is possible from such data:

3. Compare θ_{SMG} pretest with θ_{SMG} posttest for the different groups. Differences are estimates of the effect of learning.

This is a validity study on the effect of SimCityEDU with respect to a general systems modeling proficiency. We will address validity studies more broadly in the next chapter, where we will additionally suggest a further extension to investigate what Bransford & Schwartz (1999) call "preparation for future learning."

Collaboration

In some games, and thus game-based assessments, players collaborate with one another. How does this impact psychometric modeling? The answer depends on exactly what a user of the data wants to know.

One possibility is to model at the level of a team, or more generally, a collaborative unit. In this case, all of the foregoing discussion that pertained to modeling an individual's performance and capabilities applies directly to the modeling of a team as a unit. This may suffice when the team is of interest in its own right, such as when the members of an actual tank crew want to practice and improve their performance as a team. No detailing of the contributions or capabilities of individual members is provided, but discussions of team feedback and individuals' actions within scenarios can nevertheless contribute to individuals' learning.

Modeling the contributions and capabilities of individuals in collaborative units is more challenging. It is possible to have distinct models for individuals, but it must be noted how each player's actions influence the situation that other players act in. Situational features as they pertain to a given player can depend to a great extent on the actions of other players, and can differ to a great extent from one player to another. Evaluating the performance of each individual requires attending carefully to the actions of that player in light of the situations created in part by other players at any given point in time. The ideas and techniques discussed previously of identifying contingent work products apply.

Researchers and game designers have devised methods for managing collaborative action, which can be employed to sharpen evidence for the assessment aspects of GBAs (O'Neil & Chuang, 2008). Jigsaw problems provide collaborators with predetermined parts of information that is needed jointly to solve a problem, so the assessor knows a great deal about what a solution will look like and what they will have to do. Interactions among collaborators can be restricted to controlled patterns, or communications limited to a designated set of messages (Hsieh & O'Neil, 2002). Players can have assigned roles, or designated responsibilities for creating specific objects that are needed in a solution, and must have properties that allow them to interact successfully (Avouris, Dimitracopoulou, & Komis, 2003). Tasks can require non-collaborative performances as well as collaborative work, in order to distinguish players' capabilities as individuals and illuminate emergent characteristics of joint work. (Previous research suggests that people behave differently when they interact in teams than when they work alone, and team members' individual scores need not correlate highly with the team's outcome.)

One strategy that is particularly well suited to digital GBAs, and is in fact familiar and comfortable to game players, is the use of non-human characters, or avatars (Zapata-Rivera & Bauer, 2011). Avatars appear in the game environment as characters to interact with, but their behavior, while displaying some adaptivity, has known styles, knowledge bases, and behavioral patterns—all to the end of evoking targeted collaborative capabilities on the part of the human player(s) in the GBA.[13]

In collaborative problems, pertinent aspects of log file data can be considered as interacting time series for the players involved. They share paradata for situational features as covariates. The resulting multivariate time series can be analyzed with a number of modeling strategies. Few of them come from educational assessment, so they must be adapted from other fields. Strategies include dynamic factor analysis, multilevel modeling, dynamic linear models, differential equation models, nonparametric exploratory models such as social networks analysis, intra-variability models, hidden Markov models, Bayes nets, Bayesian knowledge tracing, machine learning methods, latent class analysis, neural networks, and point processes, which are stochastic processes for discrete events.

A. von Davier and Halpin (2013, in press; also see Halpin & DeBoeck, in press), for example, apply the Hawkes model, a point process model, to jointly address the capabilities of collaborating players in the case of discrete events, where what is modeled is individuals' probabilities of certain actions conditioned on previous events (see Halpin & A. von Davier, 2013, for an example with data from basketball). Among the methods they describe is an extension of IRT in which the probability of a given student's response at time t is a function of her SMV θ, but also the event history of the entire process, which includes the actions of the other individuals in the collaborative task. Independent, sequestered work following the standard local-independence IRT models is a special case against which to evaluate the degree, the nature, and the impact of collaboration. This idea can be extended to a wide variety of standard psychometric approaches that are used to model individual performance.

[13] This strategy was actually used long before digital environments were available, except the non-targeted characters were humans. The construction test for the Office of Strategic Services in World War II tasked a candidate and two assistants with constructing a simple structure within ten minutes. The two assistants played roles for each candidate: "Kippy" was passive, sluggish, and easily distracted, while "Buster" was impractical, aggressive, and critical.

Collaboration is an engaging aspect of games and learning. Capabilities that good collaboration requires are of great current interest in substantive domains. There is a large body of literature on structuring collaborative activities (e.g., Dillenbourg, 1999; Stahl, Koschmann, & Suthers, 2006; Walker, Rummel, & Koedinger, 2011). A psychometrics for collaboration, however, is only beginning. A promising route for further development will be continued development along formal modeling lines such as those in A. von Davier and Halpin (2013) and Soller and Stevens (2008), and implementation in low-stakes assessments and GBAs starting with schemas for which both design configurations and analytic methods have been worked out.

Psychometric Properties

> "Validity, reliability, comparability, and fairness are not just measurement issues, but social values that have meaning and force outside of measurement wherever evaluative judgments and decisions are made." (Messick, 1994, p. 2; emphasis original)

The terms reliability, validity, comparability, and fairness have familiar meanings in large-scale, high-stakes testing. Equally familiar statistical procedures are used to characterize these terms in these uses, and in common usage the terms are viewed as equivalent to these uses.

But Mislevy, Wilson, Ercikan, and Chudowsky (2003) argue, in the spirit of the Messick quotation above, that validity, reliability, comparability, and fairness can be viewed more broadly as qualities of assessment arguments. How they are operationalized in a given assessment situation depends on the evidence and the intended inferences a given assessment entails.

The common issue across kinds of assessments is the quality of inferences and decisions that are made from fallible and finite information. Psychometrics in general, and specifically as embodied in particular forms of these psychometric properties, is at its core about the value of information for inferences. A creative developer could certainly design a great GBA without drawing on psychometric machinery. The GBA might provide excellent evidence about students to support educational decisions. But it doesn't provide evidence about its evidence: how much, for what decisions, and how it arises from design choices, how the design choices relate back to learning. It can't challenge, it can't test, and it can't refute assumptions about evidence that are built into the GBA.

Developing this machinery and framework is useful in and of itself, regardless of how much goes into a particular GBA. The lower the stakes and the quicker the feedback cycles are, the less critical the formal psychometric machinery is. The more encompassing framework about what constitutes evidence and what are situations that can provide it is helpful nevertheless in design; this framework helps makes sure that the right learning and assessment elements are integral to game play, even if the assessment machinery per se is rudimentary. But even in this kind of GBA, having methods to characterize the value of evidence provides a metric for making improvements in the evidence as you do play testing and alpha testing. Furthermore, one needs the machinery to figure out whether the machinery is needed in a given application.

Evidence-characterizing machinery becomes increasingly important as we consider higher stakes, longer feedback cycles, or more complicated interplay among aspects of capability and aspects of actions. It is harder to sort out evidentiary relationships intuitively. A critical issue for potential higher stakes uses of GBAs – grades, badges, uses in accountability testing – is whether the evidence being obtained really supports the inferences and decisions that are being made. Without ways of characterizing the value of evidence, we don't really know how to address this question.

The nature of evidence and intended inferences in various game-based assessment use cases can be quite different than they are in large-scale, high-stakes testing, even though they embody the same underlying principles. They can appear in different forms, and become important in different ways. This section discusses how the ideas of reliability, generalizability (an extension of reliability), validity, and comparability apply to GBA.

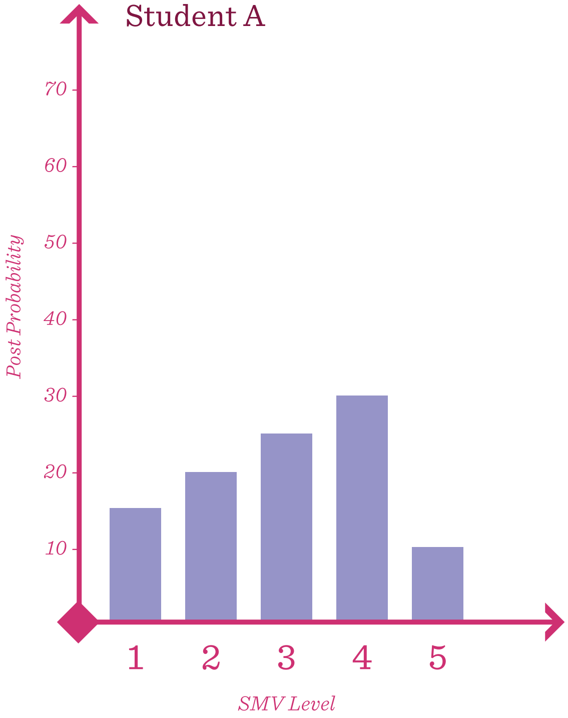
Reliability

Reliability concerns the weight of evidence in assessment data for an inference about what a student can do, understands, or has accomplished (Figure 6). Historically, in large-scale standardized tests the inference was comparing students to one another, and reliability was operationalized as by how accurately scores aligned them along the reporting scale. Although there are different ways to quantify the weight of evidence with psychometric models, two lessons from traditional reliability generally hold: More data generally provide more evidence, and data that point to different conclusions provide less evidence than data that point in a similar direction. These observations hold for GBAs as they do for traditional assessments.

For observable variables that consistent of sums or averages of similar things that all players – success in solving math problems, for example – the standard forms of calculating internal consistency reliability (KR-20, Cronbach's alpha) still work.

When different players have different amounts and different forms of evidence, as when they pursue different strategies, we can use model-based approaches to integrating evidence in terms of student-model variables (SVMs) discussed earlier, such as the SystemModeling SMV used in Jackson City. That SMV happens to be an ordered discrete variable with five levels, and what we know about a player after observing her activity is expressed as a posterior distribution over the values, given the values of the observable variables obtained from her performance – $g(\theta|x)$ in Equation 3. Figure 25 shows two students' posteriors over a five-valued ordered SMV like SystemModeling. The first panel shows belief that is more spread out across the possible values, and more concentrated in the second. (We can quantify the strength of belief by entropy: With probabilities p_k over K categories, entropy = $-\sum_k p_k \ln p_k$. Less entropy = stronger information: Entropy is highest for equal probabilities across all possibilities, and lowest when all the probability is at one possibility.

Figure 25:
Two Posterior Distributions Over a Five-Valued SMV



In this example, Level 4 is the most likely value in both situations, but there is stronger evidence in the second. There is a 35% probability that the first student is actually at level 1 or level 2, but less than 5% the second student is. The posteriors for either student could be sharpened by gathering more information, but whether we need to do so depends on the purpose at hand.

- One purpose is guiding learning in the interactive environment. The system's Activity Selection process, having this information, could adjust the situation features to just above Level 4 to both of these students. It is probably about right for the second student, but may turn out to be too hard for the first. In the dynamic environment of a GBA, though, it is easy to readjust the level down a bit or to provide support if her performance indicates she is floundering, or an engagement detector indicates she is no longer paying attention. Not a serious problem.

- Alternative purposes with somewhat higher stakes are moving on to a different challenge or assigning a grade because the goal of being at Level 3 or higher has been reached. We might be comfortable with this decision for the second student, but not the first. We would want stronger evidence for this situation.

Note though that reliability (or more generally, precision of inference) just addresses strength of evidence through the model, not whether a given strength of evidence is good enough for some particular inference or decision—and that the amount needed depends on what the inference or decision is. The lower the stakes and the easier to rectify an incorrect decision, the less evidence we need. The higher the stakes and the harder to correct in course, the stronger the evidence is needed. How much is needed for a given purpose is the province of validity, which will be addressed shortly.

Even when uses of evidence are low stakes or just internal to the game, being able to calculate reliabilities can be help one improve GBA design with regard to assessment. The reason is that psychometric indices like reliability, standard errors, entropy provide metrics for weight of evidence. We can use them in several ways, such as the following:

- To see how evidence there is in the collection of all the observables we might obtain for a given challenge – do we need to add a requirement for a pre-determined work product?
- To compare how much evidence is obtained for players who use different strategies or follow different paths through a challenge.
- To see how much evidence is added with observable variables in the form of new "detectors" constructed for patterns in log files.
- To compare different methods of combining information across features of performances in log files using A/B testing (i.e., experiments embedded in fielded games), such as which features to retain, whether to combine them with neural nets or logistic regressions or simple sums.

Measures of evidence are available for both the observable-variable and latent-variable psychometric methods described earlier. In latent variable models, the Bayesian paradigm provides posterior distributions at all times. The entropy measure mentioned above characterizes amount of evidence for categorical student model variables, and posterior standard deviations do the same for measured ones. For observable-variable accumulation, such as counts and proportions, traditional reliability indices can be used when all players (or identifiable subsets of them) encounter the same observation situations.

An approach to quantifying information that applies to both observed-variable and latent-variable methods is to divide data into parts, and use variation among the information among the parts to characterize its evidentiary value (for example, using the jackknife procedure in Mosteller & Tukey, 1977). The parts can be individual observables, different situations within challenges, or different challenges that are supposed to provide evidence about the same capability. Leaving out successive chunks of evidence, and characterizing the sensitivity of inferences to particular chunks of data, can be quite revealing, such as when certain observations have an inordinate impact on inferences. These so-called re-sampling measures of accuracy work well when model assumptions are violated, and even in many cases where there is no model at all. They work best when the chunks are similar, but can be applied nevertheless when such a partitioning is not possible to achieve. Further, they can be applied when the chunks differ in form, source, or data type, as long as there is a defined way to combine them. Indeed, this approach provides a criterion to compare alternative ways to combine disparate kinds of information.

Generalizability

As discussed above, reliability focuses on the weight of evidence from the data actually gathered from individuals. Its extension to generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), and further extension to psychometric models more generally, addresses questions of how much inferences might vary had somewhat different data been gathered.

Suppose, for example, we obtain performance data from a Jackson City player. Posterior standard deviations would tell us, in a reliability sense, how much we know about her level of systems-thinking capability as displayed in this particular game – that is, systems thinking in the context of a system of jobs and pollution in an urban situation, as simulated in the SimCity environment. But what would it tell us about what her systems thinking might have been, had the system been the water cycle, or a heating and air conditioning unit, or wolves and moose on Isle Royale, all in the SimCity environment? What if the content were the same, but it was hands-on, real-world investigation rather than SimCity, or lectures and essay tests? Generalizability helps us study how much performance varies across relevant circumstances, and thus to know how strongly performance in particular circumstances supports inferences that span the contemplated possibilities.

This is a key issue in particular for so-called 21st Century Skills like systems thinking, and others such as problem-solving, communication, and collaboration. We can surely build uses of such skills into a GBA, and obtain reliable evidence for formative feedback in this context – but as contextualized to the content and context of the game. To what degree, if any, does this evidence support inferences about other contexts and other contents, or about a more decontextualized sense of the 21st Century skill in question?

To study these questions requires we observe students engaging in multiple alternatives, or at least parts of multiple alternatives. Just having distinct samples of students take different forms tells us something about the next property, comparability, but nothing about generalizability: Even if the score distributions are identical, we need to know how much different content, contexts, and formats matter. The more any of these factors matter, the less evidentiary weight performances observed obtained under one choice of a facet provides about performance in others.

The results from generalizability studies carried out in the height of the performance assessment movement in the 1980's are sobering: The particulars of format, context, content, and occasion matter, a lot (Dunbar, Koretz, & Hoover, 1991; Linn, Baker, & Dunbar, 1991; Ruiz-Primo & Shavelson, 1996b). Yet it is precisely in in-depth, interactive, engaged experiences, with particular content in particular contexts, that students learn best, and it is such situations that education is meant to help prepare students for. In Jackson City, for example, students are faced with a complex realistic problem. They must monitor and continually adjust their strategies. They have many of the tools and information real planners have, and many of the responsibilities and constraints.

Understanding the generalizability properties of GBAs is critical for understanding how and when they can be used for different assessment use cases. The task depth and specificity that serve learning well, and are matched nicely with formative assessment uses, help students understand concepts in particular contexts (and, ideally, in ways that will help them adapt the concepts to next contexts; Bransford, Franks, Vye, & Sherwood, 1989). Extended tasks are also suited to large-scale educational surveys, where interest lies in capabilities in populations rather than in making precise inferences about individuals.

On the other hand, assessments that must support high-stakes uses for individuals and must obtain direct evidence about capabilities for acting in complicated situations usually have to observe performance in several tasks in order to overcome the low generalizability problem. For licensing physicians, for example, the National Board of Medical Examiners (NBME) currently uses 12 tasks in its computer-based patient management test and 12 simulated patient tasks in its clinical skills examination.

Comparability

High-stakes assessments are meant to accurately compare the capabilities of examinees across different times and places, for purposes that hold substantial implications such as grades, licensure, certification, employment placement, or college admission. If students are administered different forms of an assessment, considerations of fairness demand that the different forms are comparable with respect to their difficulty, the content they cover, and the accuracy of the results. Achieving comparability in the classical sense is achieved by designing equivalent challenges and imposing standard testing conditions. For example, the NBME standardized-patients exam samples carefully across factors including medical condition (e.g., cardiovascular, musculoskeletal, neurological, respiratory, etc.), age, gender, and type of physical findings. A further statistical step of test equating can be used to fine-tune the relationship between scores from different test forms (Kolen & Brennan, 1995).

For the reasons discussed above in the section on generalizability, we would not expect this degree of comparability across different GBAs. There are sources of difficulty related to background knowledge, for example, that can vary substantially from one student to the next. Even within the same GBA, different students may be following different paths, will be differently familiar/comfortable with interfaces and representations, have increased or decreased engagement due to narrative structures and game pressures.

Some of these variabilities can be managed by design and others by modeling. Domain Analysis can provide information about features of situations that demand certain kinds of knowledge and skill, and features that affect difficulty and focus. Designers can use this knowledge to craft situations that both ensure different players are challenged on the same capabilities, even if the challenges adapt to their ongoing levels of performance. Using psychometric models to synthesize evidence in terms of a common latent-variable space allows performance in situations that differ on the surface to be expressed in a common framework, if comparisons are needed. The psychometric models such as Bayes nets and structured IRT that specifically include situational features are particularly useful in this regard.

As is the case with generalizability, considerations of comparability are for GBAs vary with use cases. When comparisons among individuals are required, stricter requirements for comparability are necessary—and, with GBAs, more difficult to attain if depth, interaction, and engagement are required. When the purpose is learning, comparability remains important in that learning goals must be addressed no matter how the GBAs are adapted to different players.

Validity

Validity is paramount among psychometric principles. It speaks directly to the extent to which inferences and actions about students, based on assessment data, are justified (Cronbach, 1989; Messick, 1989). Establishing validity entails making the warrant explicit, examining the beliefs and evidence it relies on, and testing its strength and credibility. Because validity pertains to inferences and actions based on assessment information rather than assessments per se, validity investigations will take different (though overlapping) forms for different GBA use cases.

Embretson (1983) distinguishes between lines of validation that concern why data gathered in a certain way ought to provide evidence about the targeted capabilities, and lines that investigate relationships of resulting scores with other variables such as correlations with other measures or consequences of acting on the scores. These are called, respectively, arguments about "construct representation" and arguments from "nomothetic span."

For all GBA use cases, the background research in Domain Analysis grounds construct-representation evidence for validity, and the ECD framework helps make explicit how this research is embodied in the elements and the processes of an assessment. Jackson City builds on research on systems thinking. That research led to the definition of the SystemModeling student model variable (Table 2), the general design pattern for creating situations to get evidence about students' thinking about systems (Table 3), and evaluation produces for characterizing performance in the Jackson City activities (e.g., Table 4). Showing how the activities in an assessment evoke all the facets of the targeted capabilities is construct-representation evidence of validity, and it applies to all use cases. Failing to evoke some aspects of the targeted capabilities is a threat to validity that Messick (1989,

1994) called "construct under-representation." Simulations, performance assessments, and GBAs can improve construct representation in assessments by including interaction, multiple steps, a wider array of actions and representations, and more open-ended spaces for assembling and carrying out strategies.

Another threat to validity Messick identified is "construct irrelevant variance." This means knowledge or skills other than the targeted ones are required for good performance, and these demands hinder some students. Even as simulations, performance assessments, and GBAs allow for greater construct representation, they introduce more potential for construct-irrelevant variance. Lacking background knowledge, not knowing how to use the interface, and not knowing what is expected are all factors that can cause some students difficulties. Tutorials, help, and most importantly support from outside the small-g game help reduce these kinds of construct irrelevant demands.

However, the very factors that can make games engaging—narrative lines, competition, time pressure—can also work against some students. This is not a problem if using a GBA is a choice for learning and there are alternatives for students who don't like the GBA. It is a serious problem if all students are required to use them and the results are high stakes.

So construct representation issues, and the background research and ECD design strategies marshaled to address them, are shared by any purposes a GBA may be used to support. But when it comes to what Embretson (1983) called nomothetic span lines of evidence for validity, we must look more closely at the particular inferences or actions being addressed. We can see particular validation issues distinguished in various use cases.

- *Information for internal game purposes.* Information for game play includes obtaining information about a player's capabilities in order to adjust game situations and affordances (e.g., whether to "unlock" a tool). The assessment aspect of internal use is adapting difficulty or focusing evidence-gathering with respect to proficiencies. These are decisions for short term feedback loops and are generally easy to re-adjust as new information becomes available. Validity evidence can be gathered in A/B testing, where different versions of a game use different rules for adaptation, or no adaptation at all, in certain portions of play. (A/B tests are on-the-fly experiments during play, using criteria that are captured as a matter of course. Random assignment affords strong evidence for these local validity questions, while being easy to carry out and transparent to the players.) The criteria for assessing the validity of the decisions are reliability metrics for the evidence in that portion and subsequent ones, engagement and enjoyment metrics that indicate whether appropriate levels of challenge have been maintained, and learning metrics such as proficiency levels at the end of a challenge or time required to succeed in a challenge.

- *Formative assessment: Information for students.* A GBA can also provide information to a student as they play or at the end of sessions or challenges. This information, based on patterns gleaned from their play and their work products, is meant to give them better insight into their progress, and how they might enhance it. Examples of validity evidence in the small-g setting are again better performance (in comparable groups) and quicker advancement. Validity evidence outside the game could include students' use of terminology, concepts, and representations in the feedback and reports as they tackle other problems outside the game, or discuss the game and related topics with their peers.

  The goal of formative assessment is, of course, learning. Any validation of the effects of a GBA that includes formative assessment should include evidence about learning beyond the immediate context of the GBA itself. If a claim is made about the level(s) of proficiency of a student at the conclusion of play, to what degree do other contexts exhibit capabilities at comparable levels? Kinds of evidence can be distinguished by degree of transfer, as suggested in Table 7.[14] In all cases, pre-post designs provide stronger evidence about the efficacy of the formative assessment system, although they can be more difficult to carry out in practice.

- *Formative assessment: Information for teachers.* In addition to whatever information a GBA may provide to students themselves, a GBA can also provide information to teachers to support learning. This can range from digitally collected and summarized displays of students' progress, access to more detailed information about the play of individuals, to big-G activities that include, for example, lesson plans, advice for discussions, and informal assessments outside game play. There are two levels at which one can gather evidence to evaluate the effectiveness of formative assessment information with teachers as users. The first is student-level, and might be considered indirect: All of the techniques discussed above in connection with information to students to support their learning are relevant, again because student learning is the ultimate goal. The second level is more direct: In what ways do the information and affordances the GBA provides the teacher impact classroom practice? Herman, Osmundson, and Silver (2010) discuss methods for studying these issues. They underscore the need to distinguish impacts on practices and activities that teacher-level formative assessment information brings about and the quality of teacher inferences that are based on the information.

[14] As discussed in the section on multiple attempts in Chapter 11, the posterior distribution of the SystemsModeling SMV at the end of play might be thought of as a measure of proficiency in the local SimCityEDU context (θ_{SML}) or of systems modeling capabilities more generally (θ_{SMG}). These are distinguished neither by the psychometric model or the data, but by their role as evidence-summary for two different intended inferences.

Table 7:
Types of Data for Validity Studies

Degree of Transfer	Description	Example Criterion
Near (Proximal)	Ability to do similar kinds of things with same content in same context	A new but not radically different Pollution City Challenge!
Near-Medium	Ability to do unfamiliar kinds of things with same content in same context	A new and novel Pollution City challenge—e.g., new elements with different behavior in the system
Near-Medium	Ability to do familiar kinds of things with different content in same context	A challenge in the SimCity environment with a different system (e.g., invasive species in food web)
Near-Medium	Ability to do familiar kinds of things with same content in different context	A challenge like the Pollution City ones and addressing the job/pollution system, but with hands-on setting, or different simulation environment, or more traditional assessments.
Far-Medium	Ability to do unfamiliar kinds of things with different content in same context	A new and novel SimCity challenge, with a different system.
Far-Medium	Ability to do familiar kinds of things with different content in different context	A challenge like the Pollution City ones but with a different system, in a different environment (hands on, different simulation, traditional assessment)
Far-Medium	Ability to do unfamiliar kinds of things with same content in different context	A new and novel challenge addressing the job/pollution system, in a different environment (hands on, different simulation, traditional assessment)
Far (Distal), Directed	Ability to do unfamiliar kinds of things with different content in different context	A new and novel challenge addressing a new system, in a different environment (hands on, different simulation, traditional assessment)
Far (Distal), Self-directed	Occurrence of student actions applying the ideas, concepts, identities in other areas of life	Volunteer for recycling, further pursuit of topics, use of concepts & representations in different courses or life situations

- *Information for designers.* Several times we have mentioned that models and tools from psychometrics make it possible to quantify evidence about students' capabilities. This provides a metric for play testing and for further improvement from a fielded game. Improvements can take place at the level of the game experience or at the level of evidence management. Regarding the game experience, designers can modify game elements such as situations, challenges, and affordances, in order to improve information about players' capabilities without unduly degrading game play. Sources of confusion and construct-irrelevant can be identified and corrected. Additional actions or work products can be incorporated to capture more information, which if properly utilized becomes additional evidence. Regarding information management, explorations of ongoing data can be the basis of improved evidence-identification rules for existing work products, development of additional contingent work products, and discovery of additional observable variables. The fit and calibration of psychometric models can be improved. Validity evidence for this use case would consist of analyses of designer behavior. Qualitatively, do they in fact use such data to improve the assessment properties of the GBA? What kinds of activities do they employ, and how well are they integrated with ongoing game play improvements? Do they fit in with the ECD framework in ways that feed forward to new projects as well? Quantitatively, how frequent are modifications that are motivated by psychometric data, and what is their effect on reliability measures?

- *End of course assessment.* End-of-course assessment can represent for a number of use cases that attach medium-high stakes to results. A "badge" certifying successful completion of learning activities, such as in Jackson City, systems thinking at a specified level linked to a recognized standard. Here the GBA results contribute directly to a signifier of accomplishment. To validate this use case requires converging evidence about the capabilities we want students to develop by working through the game (both small g and big G). Exactly what is desired for either a grade or a badge is to be determined by the system in which it is embedded, so the options for a system might appear anywhere in the taxonomy of Table 7. Most courses, for example, look for Near and perhaps something like Near-Medium transfer. A badge system might want more.

  A particular kind of transfer inference is what Bransford and Schwatrz (1999) call "preparation for future learning" (PFL). PFL means learning in such a way that what is learned in a given local or situated context develops resources that will aid in learning in other contexts. An example is the previously-mentioned Hydrive project, in which the Air Force wanted to develop a practice system for troubleshooting the hydraulics systems of the F-15 aircraft. Their goal was more than helping trainees to troubleshoot that particular aircraft, however; they wanted the trainees to learn in ways that would help them learn faster if they were transferred to different systems (such as the more similar F-16, or the more different C-130 transport aircraft). To do this, Gitomer and Steinberg (1996) grounded the interface and feedback in the language and representations of Newell and Simon's (1972) general problem-solving framework: they used terminology and representations such as active paths, space-splitting, and serial elimination that not only described the specifics of trainees were learning, but would apply to other new systems they might encounter. To gather validity evidence for PFL, one needs to examine comparable groups on the time required and proficiency obtained in a new system, after (1) an experience meant to promote PFL, (2) an unrelated activity, to serve as a control, and (3) other learning approaches that would be of interest to compare, such as an experience meant to improve local learning but not necessarily PFL. A key point is that two instructional approaches can be comparable in local learning and even one-off assessment in other contexts, but differ with respect to PFL in new learning situations.

- *Large-scale accountability assessment.* Logistically it is possible to include one or more focused game experiences as part of state large-scale accountability tests. The arguments for doing so are the potential for increased engagement and obtaining direct evidence about interactive and constructive capabilities. High-stakes uses such as these elevate the importance of reliability and generalizability issues. Validation of this usage would include a psychometric component, namely determining the contribution of such data to the variables being measured in the assessment system, and a qualitative component, namely through observation and post-experience interviews the levels of increased engagement on the one hand, and difficulties and non-engagement on the other. In the psychometric component, key indices would be amount of non-response, difficulty parameters, and discrimination indices (low discrimination means little contribution to the intended overall measures). Particular attention would be focused on sources of construct-irrelevant variance: prior knowledge, expectations, ease of use, interaction with cultural backgrounds, confounding of game goals with evidentiary goals. The same depth of student experience that can aid learning uses of games may not serve well for this rather different purpose. It might be the case that a challenge like Jackson City does indeed provide some evidence about students' systems thinking, and this is certainly central to science learning standards. But it can also be the case that enough other factors affect performance that the information gained does not justify inclusion in a 'drop in from the sky' high stakes test. This would be the focus of the validation studies.

- *Large-scale educational surveys.* A large-scale survey—as opposed to a test—could also include a game like Jackson City to obtain information about distributions of capabilities in populations. The same qualitative considerations noted above apply, such as provoking engagement versus non-engagement and construct-irrelevant sources of variance. Psychometric considerations such as reliability is also a concern, but at the level of providing useful information about populations rather than individual students. This is a much more forgiving environment, but still the value of the information trades off against the time it uses.

Implications for Design

The view of the preceding chapters has been primarily conceptual and structural: What are key concepts in game design, assessment design, and psychometrics, and how do they interact in (primarily formative) game-based assessment? What kinds of models and processes must be implemented to make a GBA function as an assessment? This chapter looks more closely at design processes for these hybrid creatures. It draws on our experience in GlassLab designing SimCityEDU: Pollution Challenge! and other GBAs.

The following sections address the following topics in turn. First is a description of a design approach we developed in GlassLab called Evidence Centered game Design, or ECgD—a fusing of the principles of ECD assessment design framework and "best practices" in the design of recreational games. Second is a discussion of two representation forms, macrodesign documents and microdesign documents, which we have found helpful in carrying out the ECgD process. Third is our encouragement of re-usability and modular design of GBAs wherever feasible, to improve both the efficiency and the validity of GBA design.

Evidence-Centered game Design (ECgD)

Game design and assessment design are distinct domains with their own languages, their own distinct goals and constraints, and methods for tackling them. GBA is an exercise in design under constraints, with goals and constraints that come from two distinct worlds. A successful design needs to strike a good balance across domains.

Researchers have compared experts' and novices' design processes in domains such as architecture, where the artifacts are complicated and the constraints are numerous and competing (e.g., Katz, 1994). The process is inevitably iterative, for experts and novices alike—starting with rough provisional designs that address some constraints, and successively revised to meet more. The process often involves prototyping and testing. Although both novices and experts designed iteratively, experts more often needed to scrap large portions of work when they seemed to be farther along in the process. The reason is that experts addressed hard-to-meet or conflicting constraints early on in their prototypes. Novices would move ahead rapidly, better satisfying a subset of goals, but running into trouble when they tried to incorporate a hard-to-meet or conflicting constraint into their current provisional design.

The implication for the design of GBAs is address game considerations and assessment considerations jointly, if loosely, from the very start of the design process (Mislevy, Behrens, DiCerbo, Frezzo, & West, 2012; Riconscente & Vattel, 2013). In the GlassLab project, we have developed a design process called Evidence-Center game Design (ECgD) to do this.

ECgD must synthesize the two design frameworks shown in Figure 26. At the right is the ECD framework discussed extensively in the previous chapters. At the right is a version of the so-called agile design process typical of recreational video games. Agile software design processes emphasize rapid implementation and testing of successive, at the beginning simple, versions of a product. It is contrasted with a waterfall process that attempts to lay out all requirements at the beginning and create a design that meets all of them—and rarely does. Rather, in an agile process the designers expect that through rapid cycles they will come to understand, through the experiences of trying to make something work and seeing how users respond, what the requirements and constraints actually are.

The result is the more unified process suggested in Figure 27. From initial views of game, learning, and assessment perspectives, early prototypes embody some of the most important ideas of each to produce sketches of play around situations that are central to the domain and evoke evidence of players' capabilities. Successive cycles of testing, evaluating, and brainstorming refine the artifact, but designers strive to maintain this intimate synthesis of goals across domains, achieved in a unified artifact.

Figure 26: Design Frameworks for Games and Assessment That Must be Integrated

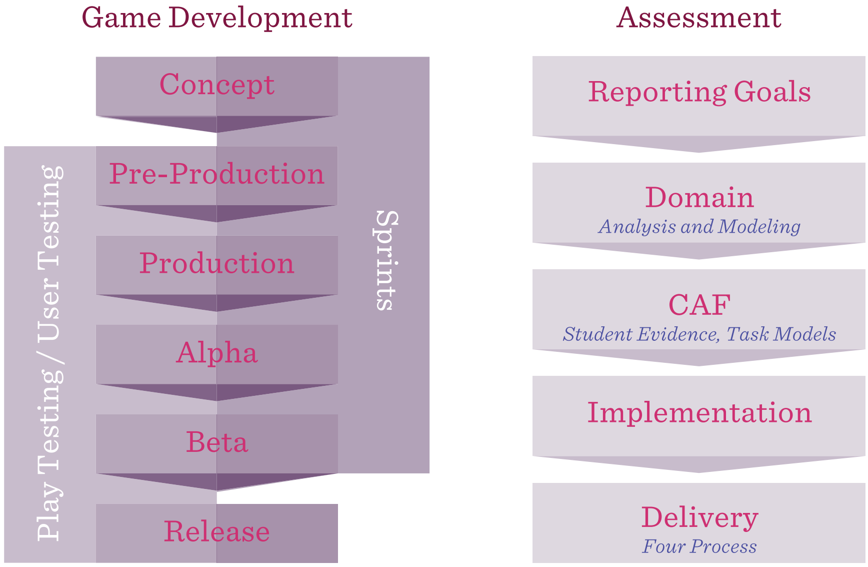
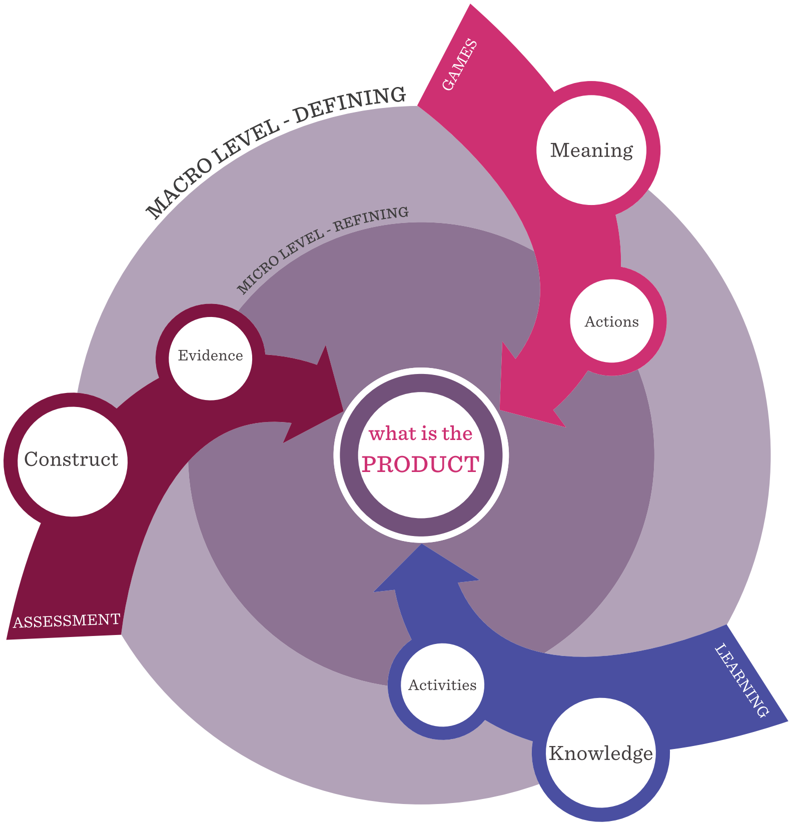


Figure 27:
Conceptions From Different Semiotic Frames Must be Synthesized in an Artifact



Evidence-Centered Game Design (ECgD), then, is a process for creating video games that can function as assessment and learning tools for competencies defined externally to the game itself. ECgD includes the following four components:

1. Definition of competencies from a non-game realm.
2. A strategy for integrating externally-defined competency with gameplay competency.
3. A system for creating formative feedback that is integral with the game experience.
4. A method for iteration of the game design for fun, engagement, and deep learning, simultaneous with iteration of the assessment model for meaning and accuracy.

Because few designers come to a GBA design as experts in all the domains that are involved, and there is not at present as well-developed practice for GBA design, everyone on a GBA design team will find all four of these components uncomfortable and unfamiliar. Both assessment designs and game designers will find ideas that are intuitive and others that are jarring, blended together. A careful definition of each helps show how they come together.

Definitions of Competencies

Both games and learning have at their heart the notion of competency. Becoming skilled at a game is a process of learning the game's mechanics, procedures, dynamics, and strategies; developing competence in substantive domains involves gaining both understanding and practical knowledge around targeted knowledge, skills, and abilities (KSAs), or what we will call here curricular competency. In this way the two are very similar.

Where ECgD creates specific demands is that in order to be considered educationally useful, the game must integrate curricular competency as a game competency. While it is conceivable to use existing games and retrofit analysis of competency (as has been done for example with Portal 2 or Minecraft), this is a different and separate process from ECgD. ECgD presumes that the effort is to unify the academically-valued competency with the gameplay.

ECgD is not retrospective analysis. That is, it is not about trying to take a previously-existing recreational game and figure out whether and how it is learning and assessment of curricular competencies. ECgD presumes that the game's mechanics and goals are made congruent with the learning goals from the beginning. While the GlassLab project believes that retrospective analysis is a promising avenue, it stands outside of the ECgD process. (SimCityEDU: Pollution Challenge! actually blends ECgD and retrospective analysis, in that it builds on a pre-existing SimCity mechanics and platform, but engineers challenges and interactions, and constrains some SimCity functionality, in order to center around, and explicitly bring out, systems thinking.)

ECgD is not gamification. If the activity of gaining competency in the game is not cognitively aligned with the activity of gaining competency in the targeted KSAs, then this is not a product of an ECgD process. Gamification focuses on engagement through game mechanics; ECgD focuses on creating game-like learning, where a central part of that learning is the curricular competencies.

Strategy for Development of Mechanics

ECgD next incorporates a process of generating game mechanics out of specified KSAs. This process is tied closely to the ECD—specifically growing from a general understanding of the kinds of situations people need to act in and the kinds of things they need to do with regard to the KSAs—as well as to canonically understood methods of developing computer games.

This process begins with identifying a practice of the expected competency. So for example in the case of argumentation, the practice is the construction of an argument to be used in a context, such as is done in the computer game Phoenix Wright. With this in hand, the game designer focuses on creating mechanics and milieu for the precise practice, within the context of an interactive computer game. At the same time, the assessment designer is defining what opportunities lie within the practice for evoking and accumulating evidence. These two processes work tightly in concert because

ECD is about evidence of reasoning, not outcome, and successful game design creates moment to moment entertainment, in other words, focused on the reasoning, not winning or losing.

A successful ECgD process proves out these steps, before building in any explicit feedback systems, either explicit or implicit. Proof, in playable software, of an integration of specific competency and fun game behavior, is considered a gate of the process.

Feedback Mechanisms

ECgD presumes that the product will make use of formative assessment for the benefit of the student, and the instructor, throughout the game experience (Shute, 2008; Shute & Kim, 2013). This formative work takes two forms:

- Enumerated, usually textual, feedback to the student and/or instructor/parent.
- Modification of the play experience according to believed competency.

In the case of enumerated feedback, generally this will be provided in-fiction (to both the student and the instructor or parent), and will include not only an assessment, but also helpful feedback to improve performance. Other subtle features such as a sense of history (you've really improved!) are also valuable tools.

The modification of the game play experience presumes an inter-relation of the game's state machine and a separate but interlocked state machine built for assessment so that evidential needs for the assessment component can be coordinated with game play. For example, the assessment state machine may inform the game state machine not to spawn a key for a door just yet, until the player has more thoroughly demonstrated the competency matched with this particular room.

Crucially, both modes of feedback are developed simultaneously and implemented in prototype form at the same time. This remains true even when properly mature assessment elements may be months away (i.e., more work products; refinements and additions to the set of observable variables; tuning, estimating, and critiquing measurement models). In its stead, placeholder assessments (sometimes even human-controlled) are used to prove out the experience.

Since the presentation of assessment is no different from the presentation of the game, this process of prototyping both presentations simultaneously is a key tenet of ECgD.

A Method of Iteration

All ECgD products are presumed to enter the marketplace "flawed," both in terms of the quality of the game, and the quality of the assessment. The core of the game experience, of its integration with competency, and its dynamic relationship with assessment, is in place and fully functional prior to the product being considered 1.0. However, assessment requires substantial numbers of players to refine and verify its models, and as refined assessment inevitably suggests modifications in the game experience, the game design is assumed to evolve and change as well, even beyond the now-familiar cycle of frequent iteration in online-hosted game experiences.

While ECgD cannot provide a Secret Sauce for this iteration process, it does require a subtle and open-minded notion both of game quality and especially of assessment validity as the game first enters the marketplace. However since ECgD is exclusively intended for use in formative assessment environments, the ability of wise teachers and instructors to correct issues (and assist the developers in correcting the product) during the early weeks and months of its launch is part of the ECgD painting.

Summary

ECgD is blend of familiar and comfortable elements and strange and counter-intuitive elements—and which are which are different for game designers and assessment designers. It is, certainly at first, an unfamiliar and uncomfortable process for all involved.

- Game designers must allow their core loops to be driven by a competency goal rather than a purely emotional inspiration.
- Technologists must integrate multiple and disparate state engines into a consolidated piece of software and experience.
- Assessment designers must allow loosely formed assessments to be integrated, made visible, tested and iterated, even in a live product in use by actual students.
- Teachers must tune their senses to the formative assessments being generated by the game, including where students may receive differentiated feedback and even differentiated gameplay experiences.

All of this said, we believe that ECgD is a key set of methods for creating games that successfully integrate and align games and learning, incorporating assessment as a shared language and shared toolkit.

Macro and Micro Design Documents

As a family of practices, over the past 15 years, ECD has used a variety of design objects to express components of an assessment design including student (aka, competency), evidence, and task models (Mislevy, Steinberg, & Almond, 2003), design patterns (Mislevy, Riconscente, & Rutstein, 2009), task templates (Riconscente et al., 2005), and augmented Q-matrices (Almond, 2010). ECgD incorporates two new design objects to coordinate and align the satisfaction of constraints drawn from games, learning, and assessment as the design work progresses. Because claims are about learning, the evidence model is embedded along with game and learning components in a macrodesign matrix that is organized around an overall instructional pattern. The detailed evidence and task designs are embedded in microdesign documents in conjunction with the detailed game and learning design. In addition, the competency model is expressed as a set of learning progressions (Corcoran, Mosher, & Rogat, 2009), such as the learning progression for systems thinking in the Jackson City mission (Table 2).

The macrodesign matrix expresses connections among the game, learning, and assessment design work and helps align this work, especially during early and middle phases of ECgD. Each row in the macrodesign matrix represents a coherent part of the educational experience while each column represents a specific game, learning, or assessment concern. For a game that is being designed as part of a larger curricular unit, the left most column defines the instructional sequence including the series of missions in the game as well as other out of game classroom experiences that connect to them. Within the SimCityEDU: Pollution Challenge! design there are several kinds of educational experiences that a part of this instructional sequence including playing a specific mission such as Jackson City (the specific challenge, or SC), using a digital learning tool connected to the challenge (LT), engaging in a classroom activity with the teacher and other students (CE) (e.g. discussion with one other student and then the whole class reflecting on the use of the causal loop diagramming tool), and interstitials (IS), or key transitions between activities that demand feedback from game or reporting infrastructure and are critical to the design. Specific instances of these activities (e.g. the five SimCityEDU: Pollution Challenge! challenges of increasing complexity that students work though) define the instructional sequence and the left most column of the macrodesign matrix. A given row then defines the game, learning, and assessment specifications that come into play for the activity at the beginning of the row.

Table 8 provides details of two of the roughly 40 rows: the row for the Jackson City challenge described earlier and a row for a learning tool experience that is intended to build upon the prior game play to support students' competency in reading text to a) gather evidence to support a claim and b) integrate meaning across texts and diagrams.

Table 8:
Two Rows from a Macrodesign Matrix

Type	Time	Activity	Learning	Student	Goals and Scaffolds	Notes	Modeling Systems L1	Modeling Systems L2	Modeling Systems L3	UI
SC	15m	Jackson City: Bivariate puzzle challenge; hard to solve without understanding the system.	Bivariate puzzle requires understanding of single causes to multiple effects, not just single chains.	"This is cool because it combines what I've learned in the past two missions... but it's harder too.	Difficult puzzle; may want to integrate causal loop as a hint in-game	This problem should be tightly time-boxed, in assumption that it may not go well for many or most students. Students should be encouraged to reboot the experience, with support between reboots.	1. Solution (unsolved). 2. Solution path: actions do not match known causal links 3. Little or no exploration of data	1. Solution with one cause to one effect only 2. Solution path - search for macro level observables 3. Solution path - exploring data for directly observable variable	1. Solution (solved) with substantial scaffolding. 2. Solution path - exploring data for directly observable variables and hidden variables	Zoning, managing power grid, data vies
LT	10m	Reading activity (based on Sim City™ experience)	ELA learning acround extracting data from text. Also includes filling in causal diagram.	"This is cool, that's the data from my gameplay yesterday."	Introduces students to the task of extracting evidence and information from a text that should resonate as very familiar		Text selected does not align with game play nor fit in diagram	Text selected fills out diagram (quantitites, dynamics) but not justification	Text selected fills out diagram (quantitites, dynamics) and justification	

The row for Jackson City includes several columns pertaining to learning. The "learning" column defines the learning goals (notion of single causes to multiple effects); the "goals and scaffolds" column includes hints to include when student show evidence of difficulty; and the "Notes" column includes other design components (in this case, time-boxing) that should be included. Several cells pertain to game design directly; the "Student" column expresses what the game design would like players to be feeling and thinking during this activity; and the UI column expresses what features of the game need to be available. Lastly, there are columns that connect most directly to assessment: the columns labeled "Modeling Systems L1" (for level 1), L2, and L3 include the expectation for what evidence can gather for the different levels of the learning progression that is the focus of the assessment (systems and system modeling).

Overall the macrodesign includes and expands aspects of a Q-matrix in that the activity rows and Modeling Systems columns define an item-by-competency matrix; however, these are expanded in the macrodesign to include each level of competency with each cell containing a description of evidential expectations rather than a binary yes/no. The macrodesign also incorporates aspects of what game designers call an experience matrix, which includes the major game flows, mechanics, and expectations about players' experiences with the game. The Q-matrix and the experience matrix guide the early and middle phases of assessment and game design respectively. The macrodesign within ECgD combines aspects of both matricies and incorporates learning expectations to guide and coordinate the games, learning, and assessment design.

While the macrodesign matrix was organized by curricular sequence for SimCityEDU: Pollution Challenge!, there are other ways of organizing it depending upon the design purpose and needs. In other current work that is focused on game-based formative assessment for informal learning (e.g. kids playing at home, in the car, etc.) and where there is no in-class activity or curricular sequence, but instead more seamless game play, we've found it more useful to directly organize game flow around the sequence of learning goals and assessment needs that support those goals, and then structure the macrodesign matrix by the resulting game loops and mechanics. In both projects, the intent is to organize the overall design into more manageable, coherent modular units with respect to game flow, instructional sequence, and evidence collection.

The microdesign is a more detailed specification – a playing out of the details for a row of the macrodesign. Each row in the macrodesign has a corresponding microdesign. For example, the microdesign for Jackson City is a roughly 8 page document that formed the basis for implementing that specific challenge. It describes the details needed for creating the challenge within SimCityEDU: Pollution Challenge!, including the purpose of the challenge, the information that needs to be provided to students at the beginning of the challenge, the technology features, graphics and reporting requirements, full description of the kinds of evidence to be collected, needed features of the challenge, discussion questions for teachers to use, and detailed play sequence including fail states and feedback. While the intent of this document is to provide enough information for an initial implementation of the row (in this case a particular game challenge), it does not include the full set of scoring rules and measurement models. These are described in a later design document in a later phase after a moderately small pilot, a "mini-tryout," is conducted. This user testing provides an early empirical basis to guide designers' judgments to define provisional scoring rules and set initial parameters in measurement models. Even at that phase, the rules and models are still considered theories for which further evidence of validity will be collected in later alpha and beta trials.

Modularity and Re-usability

By this point it is clear that there is a lot going on in designing a game-based assessment, cutting across several domains of knowledge and design. It is hard to do, especially at first. And once designers figure out how to do something, it would be nice for them and for other designers to not have to rediscover it again the hard way. We have an incentive to devise ways to represent what we learn in representations that can be re-used—whether conceptual or mechanical. This should help with efficiency especially in GBAs, because it can reduce the hardest problem, namely finding clusters of design elements that together address joint cross-domain design goals and offer integrated solutions. The more of these we find and the better we can encapsulate them for future projects and other designers, the better the chances they can leverage these elements or modify them to solve some recurring kind of problem. These are not new ideas. At the conceptual level, Christopher Alexander (Alexander, Ishikawa, & Silverstein, 1977) introduced design patterns in architecture to describe recurring problems in a general form, lay out strategies for tackling them, and providing talked-through examples. Gamma, Helm, Johnson, and Vlissides (1994) brought the approach to software engineering.

One way to facilitate modularity and re-usability is to have a design framework which, while flexible, provides a common way of thinking and talking about problems, and common representational forms for expressing solutions. This is a central motivating goal of ECD, of course. Representations for recurring argument elements have been developed in the Domain Modeling layer; for schematic elements in the CAF; and mechanical elements and procedures in the Delivery layer.

In Domain Modeling, the Principled Assessment Design for Inquiry (PADI) project developed design patterns to help test developers target hard-to-assess aspects of science such as systems thinking (Cheng, Ructtinger, Fujii, & Mislevy, 2010) and model-based reasoning (Mislevy, Riconscente, & Rutstein, 2009). The PADI assessment design patterns include attributes such as characteristic and variable features of tasks that can evoke evidence about targeted areas of competence (they focused on inquiry in STEM), and potential work products, observable variables, and evaluation procedures. One could extend the assessment design pattern form to include compatible game-design elements, such as game mechanics that might be well-suited play that also provided assessment information about the targeted competencies.

For example, the mechanic of a teachable agent (Biswas, Katzlberger, Bransford, & Schwartz, 2001) serves well for knowledge and skills that can be expressed with a symbol-system representation. The agent would collaborate with the player, and could do things in the game space the player couldn't—but only if the player taught the agent how to act or reason to solve problems using the targeted competencies. In this way, the player must come to understand the procedures well enough to get the agent to solve the kinds of problems in the domain of interest.

Many game editing programs are available for various aspects of assessment development, to help designers focus their energies on content and interaction rather than low-level programming. Just as game mechanics are available for reuse in environments that may look quite different on the surface, so too are forms of work products and evaluation strategies for assessment (e.g., Scalise & Gifford, 2006) and statistical model building blocks (e.g., Netica[15]). These can be pressed into service when games are used as assessments. Designers who know what styles of interaction support efficient evaluation can use them early on, rather than finding out down the line that the styles they happened to use did not produce good evidence. Design patterns, editing environments, and reusable objects are available and familiar to practitioners in the domains of games and assessments. These tools have lessons from experience and design strategies built into them for tackling constraints in a given domain. The need to deal jointly with constraints across domains can be supported by hybrid approaches, such as Vendlinski, Baker, and Niemi's (2008) (conceptual level) templates and (implementation level) objects for authoring simulation-based problem-solving assessments. Similarly, Mislevy, Steinberg, Breyer, Johnson, and Almond (2002) provided schemas for recurring situations around which task authors could write unique problem-solving cases for dental hygiene students, in forms that linked to re-usable task-scoring and test-scoring machinery.

[15] www.norsys.com. Downloaded May 1, 2011.

Conclusion

The initial proposal of GlassLab aimed to "design, develop, disseminate, and provide opportunities for research around new models of formative, adaptive assessment with digital games at their core. [GlassLab] addresses a need to effectively assess, in an integrated fashion, common core/domain-based knowledge and skills, as well as competencies like problem solving, systemic reasoning, and knowing how to learn (metacognition)."

These aims implicitly call for a principled approach to assessment in game contexts. But although the literature contains some good work on assessment in games (e.g., Ifenthaler, Eseryel, & Ge, 2012), it is but in its initial stages. This presentation seeks to contribute to an integrated framework for designing, implementing, and using game-based assessments; one which builds on current best practices in learning, game design, and assessment design.

The focus is how one can apply the concepts and the methods of psychometrics to this end, and we do indeed address this question. It is our belief, however, that in order to apply psychometric methodology most effectively, it cannot be done by building what one hopes is a great GBA, then "throw it over the wall" to the psychometricians, to see if they can "figure out how to score it." We think this approach is bound to disappoint. More promise is building into a GBA from the very beginning not only the elements of good learning and engaging game play, but good assessment as well for the purpose(s) the assessment aspects of the GBA are meant to serve. To develop such a framework, we have drawn on recent work not only in game design, but in measurement modeling, knowledge-based model construction, and educational data mining; and we have suggested a design methodology, called evidence-centered game design, or ECgD, that helps coordinate the different perspective which need to be brought to bear to design an effective game-based assessment.

GBA is an exciting opportunity for psychometrics—and a crucial one for the profession. There are issues of reliability, validity, comparability, and fairness, all long-standing psychometric values, which the field has developed insights and methods to address in familiar kinds of assessments. They may need to be extended, augmented, and reconceived to play analogous roles in GBA. Without the insights of psychometrics, GBA would proceed nevertheless -- and designers would need to rediscover these principles and figure out how to address them anew

Index

References

Achieve, Inc. (2013). *Next Generation Science Standards*. Washington, D.C.: Achieve, Inc.

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21(1), 1-23.

Alexander, C., Ishikawa, S., & Silverstein, M. (1977). *A pattern language: Towns, buildings, construction.* New York: Oxford University Press.

Almond, R. G. (2010). I can name that Bayesian network in two matrixes. *International Journal of Approximate Reasoning*, 51, 167-178.

Almond, R.G., Dibello, L., Jenkins, F., Mislevy, R.J., Senturk, D., Steinberg, L.S. and Yan, D. (2001). Jaakkola & Richardson (eds.), *Models for conditional probability tables in educational assessment: Artificial Intelligence and Statistics 2001* (pp. 137–143). San Francisco: Morgan Kaufmann.

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J. D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement*, 44, 341-359.

Almond, R.G., & Mislevy, R.J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement*, 23, 223-237.

Almond, R. G., Mulder, J., Hemat, L. A., & Yan, D. (2006). *Models for local dependence among observable outcome variables.* Technical report RR-06-36, Educational Testing Service. Available at: http://www.ets.org/research/researcher/RR-06-36.html

Almond, R.G., Steinberg, L.S., & Mislevy, R.J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment*, 1(5). http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2005). *A taxonomy for learning, teaching, and assessing*. Longman.

Arndt, H. (2006). Enhancing system thinking in education using system dynamics. *Simulation*, 82, 795-806.

Avouris, N., Dimitracopoulou, A., & Komis, V. (2003) On analysis of collaborative problem solving: an object-oriented approach. *Computers in Human Behavior*, 19, 147-167.

Bagley, E., & Shaffer, D.W. (2009). When people get in the way: Promoting civic thinking through epistemic gameplay. *International Journal of Gaming and Computer-mediated Simulations*, 1, 36–52.

Baker, R.S.J.d.. Corbett, A.T., & Aleven, V. (2008). More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing. *Paper 6*. Pittsburgh: Human-Computer Interaction Institute, Carnegie-Mellon University. Available online at http://repository.cmu.edu/hcii/6

Baker, R.S.J.d., D'Mello, S.K., Ma.Mercedes, T.R., & Graesser, A.C. (2010) Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68, 223-241.

Bartholomew, D.J. & Knott, M. (1999). *Latent factor models and factor analysis (2nd edition)*. Kendall's Library of Statistics, 7. NY, NY: Oxford University Press.

Baxter, G.P., & Shavelson, R.J. (1994). Science performance assessments: Benchmarks and surrogates. *International Journal of Educational Research*, 21, 279-298.

Behrens. J. T., DiCerbo, K. E., Yel, N. & Levy, R. (2012). *Exploratory data analysis*. In I. B. Weiner, J. A. Schinka, & W. F. Velicer (Eds.) Handbook of Psychology: Research Methods in Psychology, 2nd ed (pp. 34-70). New York: Wiley.

Behrens, J.T., Mislevy, R.J., DiCerbo, K.E., & Levy, R. (2012). An evidence centered design for learning and assessment in the digital world. In M.C. Mayrath, J. Clarke-Midura, & D. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 13-54). Charlotte, NC: Information Age.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *The Journal of technology, learning and assessment*, 2(3).

Béland, A., & Mislevy, R.J. (1996). Probability-based inference in a domain of proportional reasoning tasks. *Journal of Educational Measurement*, 33, 3-27.

Bennett, R.E., & Bejar, I.I. (1998). Validity and automated scoring: It's not only the scoring. Educational Measurement: Issues and Practice, 17(4), 9-17.

Biswas, G., Katzlberger, T., Bransford, J., & Schwartz, D. (2001). Extending intelligent learning environments with teachable agents to enhance learning. In J. D. Moore, C. L. Redfield, & W. L. Johnson (Eds.), Artificial Intelligence for Education 01 (pp. 389-397). Amsterdam: IOS Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice*, 5(1), 7–73.

Bransford, J. D. & Schwartz, D. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education: Vol. 24* (pp. 61-100). Washington, DC: American Educational Research Association.

Bransford, J. D., Franks, J. J., Vye, N. J. & Sherwood, R. D. (1989). New approaches to instruction: Because wisdom can't be told. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 470–497). New York, NY: Cambridge University Press.

Brown, N.J.S. (2005). *The multidimensional measure of conceptual complexity* (Tech. Rep. No. 2005-04-01). Berkeley, California: University of California, BEAR CENTER.

Brown, N.J.S., & Wilson, M. (2011). A model of cognition: The missing cornerstone in assessment. *Educational Psychology Review*, 23, 221-234.

Campione, J. C., & Brown, A. L. (1987). Dynamic assessment: An interactional approach to evaluating learning potential. In C.S. Lidz (Ed), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 82-115). New York, NY, US: Guilford Press.

Cheng, B. H., Ructtinger, L., Fujii, R., & Mislevy, R. (2010). Assessing Systems Thinking and Complexity in Science (*Large-Scale Assessment Technical Report 7*). Menlo Park, CA: SRI International. Available online at http://ecd.sri.com/downloads/ECD_TR7_Systems_Thinking_FL.pdf

Chung, G.K.W.K., Baker, E.L., Delacruz, G.C., Bewley, W.L., Elmore, J., And Seely, B. (2008). A computational approach to authoring problem-solving assessments. In E.L. Baker, J. Dickieson, W. Wulfeck, and H.F. O'Neil (Eds.), *Assessment Of Problem Solving Using Simulations* (pp. 289–307). Mahwah, NJ: Erlbaum.

Chung, G. K. W. K., Baker, E. L., Vendlinski, T. P., Buschang, R. E., Delacruz, G. C., Michiuye, J. K., & Bittick, S. J. (2010, April). Testing instructional design variations in a prototype math game. In R. Atkinson (Chair), *Current perspectives from three national R&D centers focused on game-based learning: Issues in learning, instruction, assessment, and game design.* Structured poster session at the annual meeting of the American Educational Research Association, Denver, CO.

Clark, J. S. (2005). Why environmental scientists are becoming Bayesians. *Ecology letters*, 8(1), 2-14.

Collins, A. & Ferguson, (1993). Epistemic forms and epistemic games: Structures and strategies to guide inquiry. *Educational Psychologist*, 20(1), 25-42.

Common Core State Standards Initiative. (2010a). *Common Core State Standards for English Language arts & literacy in history/social studies, science, and technical subjects*. Retrieved from http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf

Common Core State Standards Initiative. (2010b). *Common Core State Standards for mathematics*. Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

Corbett, A.T., & Anderson, J.R. (1995) Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4, 253-278.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science*. CPRE Research Report # RR-63. Philadelpia: Consortium for Policy Research in Education.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Dayton, C. M. (Ed.). (1998). *Latent class scaling analysis* (No. 126). Thousand Oaks, CA:Sage.

Deane, P. (2006). Strategies for evidence identification through linguistic assessment of textual responses. In D. M. Williamson, R. J. Mislevy, and I. I. Bejar (Eds.) *Automated scoring of complex performance in computer based testing*. Mahwah, NJ: Erlbaum Associates.

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

DiCerbo, K. E., & Behrens, J. T. (2012). Implications of the digital ocean on current and future assessment. In R. Lissitz & H. Jiao (Eds.) *Computers and Their Impact on State Assessment: Recent History and Predictions for the Future*. (pp. 273-306). Charlotte, NC: Information Age Publishing.

DiCerbo, K. E. & Kidwai, K. (2013). Detecting Player Goals from Game Log Files. Poster presented at the Sixth International Conference on Educational Data Mining, Memphis, TN.

Dillon, G. F., Clyman, S. G., Clauser, B. E., & Margolis, M. J. (2002). The introduction of computer-based case simulations into the United States medical licensing examination. *Academic Medicine*, 77(10), S94-S96.

Dillenbourg, P. (Ed) (1999). *Collaborative learning: cognitive and computation approaches* (2nd Ed). Elsevier, Amsterdam: Emerald Group Publishing Limited.

Dunbar, S.B., Koretz, D.M., & Hoover, H.D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, 4, 289-303.

El-Nasr, M. S., Drachen, A., & Canossa, A. (2013). *Game analytics: Maximizing the value of player data*. London: Springer.

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

Embretson, S.E. (1990). Diagnostic testing by measuring learning processes: Psychometric considerations for dynamic testing. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (pp. 453-486). Mahwah, NJ: Lawrence Erlbaum.

Embretson, S.E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3, 380-396.

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

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994). *Design patterns*. Reading, MA: Addison-Wesley.

Gee, J. P. (1992). *The social mind: Language, ideology, and social practice*. New York: Bergin & Garvey.

Gee, J.P. (2007). What video games have to teach us about learning and literacy (2nd ed.). New York: Palgrave.

Gee, J.P. (2008). Learning and games. In K. Salen (ed.), *The ecology of games: Connecting youth, games, and learning* (pp. 21–40). Cambridge, MA: MIT Press.

Gee, J. P. (2013). *An introduction to discourse analysis: Theory and method*. London: Routledge.

Geerlings, H., Glas, C. A., & van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika*, 76(2), 337-359.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2004). *Bayesian data analysis* (second edition). London: Chapman & Hall.

Gierl, M. J., & Haladyna, T. M. (Eds.). (2012). *Automatic item generation: Theory and practice*. Routledge.

Gierl, M.J., & Lai, H. (2012). The role of item models in automatic item generation. *International Journal of Testing*, 12, 273-298.

Gitomer, D.H., Steinberg, L.S., & Mislevy, R.J. (1995). Diagnostic assessment of trouble-shooting skill in an intelligent tutoring system. In P. Nichols, S. Chipman, & R. Brennan (Eds.), Cognitively diagnostic assessment (pp. 73-101). Hillsdale, NJ: Erlbaum.

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement*, 27, 247-261.

Gobert, J.D., Sao Pedro, M., Baker, R. S. J. D., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining*, 5, 153-185.

Goldstone, R. R., & Wilensky, U. (2008). Promoting transfer by grounding complex systems principles. *Journal of the Learning Sciences*, 17, 465-516.

Gong, S., Ng, J., & Sherrah, J. (2002). On the semantics of visual behaviour, structured events and trajectories of human action. *Image Vision Computing*, 20, 873–888.

Greeno, J. G. (1998). The situativity of knowing, learning, and research. *American Psychologist*, 53, 5–26.

Gulliksen, H. (1950/1987). Theory of mental tests. New York: Wiley. Reprint, Hillsdale, NJ: Erlbaum.

Halpin, P.F., & De Boeck, P. (in press). Modeling dyadic interaction with Hawkes process. *Psychometrika*.

Halpin, P., F. & von Davier, A. A. (2013, May). *Evaluating the roles of individual team members in team interactions*. Paper presented at an invited symposium at the annual meeting of the National Council of Measurement in Education, San Francisco, CA.

Hammer, D., Elby, A., Scherr, R. E., & Redish, E. F. (2005). Resources, framing, and transfer. In J. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 89-120). Greenwich, CT: Information Age Publishing.

Heritage, M. (2010). *Formative assessment: Making it happen in the classroom*. Thousand Oaks, CA: Corwin Press.

Herman, J., L., Osmundson, E., & Silver, D. (2010). *Capturing quality in formative assessment practice: Measurement challenges*. (CRESST Report 770). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Holland, P. W. (1994). Measurements or contests? Comments on Zwick, Bond and Allen/Donoghue. *In Proceedings of the Social Statistics Section of the American Statistical Association*, 27–29. Alexandria, VA: American Statistical Association.

Hsieh, I.-L., & O'Neil, H. F., Jr. (2002). Types of feedback in a computer-based collaborative problem solving group task. *Computers in Human Behavior*, 18, 699-715.

Ifenthaler, D., Eseryel, D., & Ge, X. (eds.) (2012). *Assessment in game-based learning: Foundations, innovations, and perspectives*. New York: Springer.

Irvine, S. H. (in press). Tests for recruitment across cultures: a tactical psychometric handbook. IOS Press, Amsterdam, The Netherlands.

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

Ito, M. (2009). *Engineering play: A cultural history of children's software*. Cambridge, MA: MIT Press.

Jones, L.V. and Olkin, I. (Eds) (2004). *The Nation's Report Card: Evolution and perspectives*. Bloomington: Phi Delta Kappa Educational Foundation, American Educational Research Association.

Katz, I.R. (1994). Coping with the complexity of design: Avoiding conflicts and prioritizing constraints. In A. Ram, N. Nersessian, & M. Recker (eds.), *Proceedings of the Sixteenth Annual Meeting of the Cognitive Science Society* (pp. 485-489). Mahwah, NJ: Erlbaum.

Kimball, R. (1982). A self-improving tutor for symbolic integration. In D. Sleeman & J.S. Brown (Eds.), Intelligent tutoring systems (pp. 283-307). London: Academic Press.

Klopfer, E., & Osterweil, S. & Salen, S. (2009). *Moving learning games forward*. Cambridge, MA: The Education Arcade: Massachusetts Institute of Technology. Available online at http://education.mit.edu/papers/MovingLearningGamesForward_EdArcade.pdf

Koedinger, K. R., Aleven, V. A. W. M. M., & Heffernan, N. (2003). Toward a rapid development environment for Cognitive Tutors. In U. Hoppe, F. Verdejo, & J. Kay (Eds.), *Artificial Intelligence in Education: Shaping the Future of Learning through Intelligent Technologies, Proceedings of AI-ED* (pp. 455-457). Amsterdam: IOS Press.

Koenig, A. D., Lee, J. J., Iseli, M., & Wainess, R. (2010). *A conceptual framework for assessing performance in games and simulation.* (CRESST Report 771). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kolen, M.J., & Brennan, R.L. (1995). *Test equating: Methods and practices*. New York: Springer.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25, 259-284.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.

Levy, R. (2006). Posterior predictive model checking for multidimensionality in item response theory and Bayesian networks. Doctoral dissertation, University of Maryland at College Park.

Levy, R. (2012, May). Psychometric advances, opportunities, and challenges for simulation-based assessment. Princeton, NJ: K-12 Center at ETS. Retrieved from http://www.k12center.org/rsc/pdf/session2-levy-paper-tea2012.pdf

Levy, R. (in press). Dynamic Bayesian network modeling of game based diagnostic assessments. CRESST Research Report. Los Angeles: National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.

Lewis, C. (1986). *Test theory and Psychometrika*: The past twenty-five years. Psychometrika, 51, 11-22.

Lewis, C. (2001). Expected response functions. In A. Boomsma, M.A.J. van Duijn, & T.A.B. Snijders (Eds.), *Essays on item response theory* (pp. 163-171). New York: Springer.

Lindley, D.V., & Novick, M.R. (1981). The role of exchangeability of inference. *Annals of Statistics*, 9, 45-58.

Linn, R.L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher*, 23(9), 4-14.

Lord, F.M., & Novick, M.R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Luecht, R. M. (2003). Multistage complexity in language proficiency assessment: A framework for aligning theoretical perspectives, test development, and psychometrics. *Foreign Language Annals*, 36, 527–535.

Luecht, R.M. (2009). Adaptive computer-based tasks under an assessment engineering paradigm. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved May 30, 2011 from www.psych.umn.edu/psylabs/CATCentral/

Margolis, M. J., & Clauser, B. E. (2006). A regression-based procedure for automated scoring of a complex medical performance assessment. In. D.W. Williamson, I. I. Bejar, & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 123-168). Mahwah, NJ: Lawrence Erlbaum.

Mislevy, R.J. (in press). Missing responses in item response theory. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory, 2nd Edition*, Volume 2, Chapman & Hall/CRC Press.

Mislevy, R.J. (2013). Evidence-centered design for simulation-based assessment. *Military Medicine* (special issue on simulation, H. O'Neil, Ed.), 178, 107-114.

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where the numbers come from. *In Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 437-446). Morgan Kaufmann Publishers Inc.

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). New York: Springer.

Mislevy, R.J., Behrens, J.T., DiCerbo, K., & Levy, R. (2012). Design and discovery in educational assessment: Evidence centered design, psychometrics, and data mining. *Journal of Educational Data Mining*, 4, 11-48. Available online at http://www.educationaldatamining.org/JEDM/images/articles/vol4/issue1/MislevyEtAlVol4Issue1P11_48.pdf

Mislevy, R.J., & Gitomer, D.H. (1996). The role of probability-based inference in an intelligent tutoring system. User-Modeling and User-Adapted Interaction, 5, 253-282.

Mislevy, R.J., & Riconscente, M.M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 61-90). Mahwah, NJ: Erlbaum

Mislevy, R.J., Riconscente, M.M., & Rutstein, D.W. (2009). Design patterns for assessing model-based reasoning *(PADI-Large Systems Technical Report 6)*. Menlo Park, CA: SRI International. . Available online at http://ecd.sri.com/downloads/ECD_TR6_Model-Based_Reasoning.pdf

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.A. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-67.

Mislevy, R.J., Steinberg, L.S., Breyer, F.J., Johnson, L., & Almond, R.A. (2002). Making sense of data from complex assessments. *Applied Measurement in Education*, 15, 363-378.

Mislevy, R.J., Wilson, M.R., Ercikan, K., & Chudowsky, N. (2003). Psychometric principles in student assessment. In T. Kellaghan & D. Stufflebeam (Eds.), *International Handbook of Educational Evaluation* (pp. 489-531). Dordrecht, the Netherlands: Kluwer Academic Press.

Mood, Alexander McFarlane & Specht, R. D. (1954). *Gaming as a Technique of Analysis.* Santa Monica, CA: RAND Corporation.

Mosteller, F., & Tukey, J.W. (1977). *Data analysis and Regression: A second course in statistics*. Reading, MA: Addison-Wesley.

Mosteller, F. & Wallace, D. L. (1963). Inference in an authorship problem. *Journal of the American Statistical Association*, 58, 275-309.

Moustaki, I. & Knott, M. (2000) Generalised latent trait models. *Psychometrika*, 65, 391-411.

National Science Teachers Association. (2012). *Next generation science standards*. Retrieved 9/16/2013 from http://23.23.182.104/access-standards/

Newell, A., & Simon, H.A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

NGSS Lead States (2013). *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press.

O'Neil, H. F., Jr., & Chuang, S. H. (2008). Measuring collaborative problem solving in lowstakes tests. In E. L. Baker, J. Dickieson, W. Wulfeck, & H. F.O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 177-199). Mahwah, NJ: Lawrence Erlbaum Associates.

Plass, J. L., Homer, B. D., Kinzer, C., Frye, J., & Perlin, K. (2011). Learning mechanics and assessment mechanics for games for learning. *G4LI White Paper #1.* New York, NY: Institute for Games for Learning, New York University. Available online at http://createx.alt.ed.nyu.edu/classes/2505/reading/Plass%20et%20al%20LAMechanics%202505.pdf

Poehner, M.E. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting second language development*. Berlin: Springer Publishing.Polti, G. P. (1868/1977). The thirty-six dramatic situations. L. Ray (Translator). Boston: The Writers, Inc.

Reckase, M.D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Richmond, B., & Peterson, S. (2001). *An introduction to systems thinking*. High Performance Systems., Incorporated.

Riconscente, M., Mislevy, R., Hamel, L., & PADI Research Group (2005). *An introduction to PADI task templates (PADI Technical Report 3)*. Menlo Park, CA: SRI International. Available online at http://padi.sri.com/downloads/TR3_Templates.pdf

Riconscente, M. M., & Vattel, L. (2013, April). Extending ECD to the design of learning experiences. In M. M. Riconscente (Chair), ECD from A to Z: Applying evidence-centered design across the assessment continuum. Invited session presented at the National Council on Measurement in Education, San Francisco, CA.

Robinett, Warren (2005). *Adventure as a Video Game: Adventure for the Atari 2600.* In K. Salen and E. Zimmerman, The Game Design Reader (pp. 690-713). Cambridge, MA: MIT Press.

Romero, C., Ventura, S., Pechenizkiy, M., AND Baker, R.S.J.D. (Eds.). 2011. Handbook of Educational Data Mining. CRC Press, Boca Raton, FL.

Roschelle, J. (1997). Designing for cognitive communication: Epistemic fidelity or mediating collaborative inquiry? In. D. L. Day & D. K. Kovacs (Eds.), *Computers, communication, and mental models*. Bristol, PA: Taylor & Francis.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996a). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching*, 33, 569-600.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996b). Rhetoric and reality in science performance assessments: An update. *Journal of Research in Science Teaching*, 33, 1045-1063.

Rupp, A. A. (2002). Feature selection for choosing and assembling measurement models: A building-block-based organization. *International Journal of Testing*, 2, 311–360.

Rupp, A.A., Nugent, R., & Nelson, B. (2012). Evidence-centered design for diagnostic assessment within digital learning environments: Integrating modern psychometrics and educational data mining. *Journal of Educational Data Mining, 4*(1), 1-10. Available online at http://www.educationaldatamining.org/JEDM/images/articles/vol4/issue1/IntroVol4Issue1P1_10.pdf

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

Sadler, T., Barab, S., & Scott, B (2007). What do students gain by engaging in socioscientific inquiry? *Research in Science Education 37*(4): 371–391.

Salen, K., & Zimmerman, E. (2004). *Rules of play: Game design fundamentals*. Cambridge: MIT.

Sao Pedro, M., Baker, R., Gobert, J., Montalvo, O., & Nakama, A. (2011). Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. *User Modeling and User-Adapted Interaction,* 23, 1-39.

Scalise, K. (2013). *Multiple grain sizes of inference in innovative assessments: mIRT-bayes as a hybrid measurement model.* Seminar for the Center for Educational Assessment, University of Massachusetts, Amherst.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment*, 4(6), Retrieved July 16, 2013, from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1653

Schmit, M. J., & Ryan, A. M. (1992). Test-taking dispositions: A missing link? *Journal of Applied Psychology*, 77(5), 629.

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

Segall, D.O. (2010). Principles of multidimensional adaptive testing. In W.J. van der Linden & C.A.W. Glas (eds.), *Elements of adaptive testing* (pp. 57-75). , New York: Springer.

Shaffer, D.W. (2006). Epistemic frames for epistemic games. Computers and Education, 46(3), 223–234.

Shaffer, D. W. (2007). *How computer games help children learn*. New York: Palgrave.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78,153-189.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias, & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503-524). Charlotte, NC: Information Age Publishers.

Shute, V. J., & Kim, Y. J. (2013). 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 Edition). New York, NY: Lawrence Erlbaum Associates, Taylor & Francis Group.

Shute, V. J., & Psotka, J. (1996). Intelligent tutoring systems: Past, present, and future. In D. Jonassen (Ed.), Handbook of research for educational communications and technology (pp. 570-600). New York, NY: Macmillan.

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. Cody, & P. Vorder (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). Mahwah, NJ: Routledge, Taylor and Francis.

Sikorski, T., & Hammer, D. (2010). A critique of how learning progressions research conceptualizes sophistication and progress. In K. Gomez, L. Lyons, & J. Radinsky (Eds.) *Learning in the Disciplines: Proceedings of the 2010 International Conference of the Learning Sciences* (pp. 277-284). Chicago, IL: ISLS.

Soller, A., & Stevens, R. (2008). Applications of stochastic analyses for collaborative learning and cognitive assessment. In G. R. Hancock & K. M. Samuelson (Eds.), *Advances in latent variable mixture models* (pp. 109–111). Charlotte, NC: Information Age Publishing.

Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L., & Cowell, R.G. (1993). Bayesian analysis in expert systems. Statistical Science, 8, 219-283.

Stahl, G., Koschmann, T., & Suthers, D. (2006). Computer-supported collaborative learning: An historical perspective. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp.409-426). Cambridge, U.K.: Cambridge University Press.

Steinberg, L.S., & Gitomer, D.G. (1996). Intelligent tutoring and assessment built on an understanding of a technical problem-solving task. *Instructional Science*, 24, 223-258.

Stevens, R., & Casillas, A. (2006). Artificial neural networks. In D.M. Williamson, R.J. Mislevy, & I.I. Bejar (Eds.), *Automated scoring of complex tasks in computer based testing* (pp. 259-312). Mahwah, NJ: Erlbaum Associates.

Sundre, D. L., & Wise, S. L. (2003). Motivation filtering: An exploration of the impact of low examinee motivation on the psychometric quality of tests. *Annual meeting of the National Council on Measurement in Education, Chicago, IL*. Retrieved from http://www.jmu.edu/assessment/wm_library/Filter.pdf

Suppes, P., & Morningstar, M. (1972). Computer-assisted instruction at Stanford, 1966-68: Data, models, and evaluation of the arithmetic programs. New York: Academic Press.

Svarovsky, G. N., & Shaffer, D. W. (2007). SodaConstructing knowledge through exploratoids. *Journal of Research in Science Teaching*, 44(1), 133-153.

Ting, C.-Y., Phon-Amnuaisuk, S., & Chong, Y.-K. (2008). Modeling and intervening across time in scientific inquiry exploratory learning environment. *Educational Technology & Society*, 11, 239–258.

Torres, R., & Wolozin, L. (2011). *Quest to Learn: Developing the school for digital kids*. MIT Press.

VanLehn, K. (2008). Intelligent tutoring systems for continuous, embedded assessment. In C. Dwyer (Ed.) The future of assessment: Shaping teaching and learning. Mahwah, NJ: Erbaum. pp. 113-138

Vendlinski, T. P., Baker, E. L. & Niemi, D. (2008). Templates and objects in authoring problem solving assessments. In E. L. Baker, J. Dickieson, W. Wulfeck & H. F. O'Neil (Eds.), *Assessment of problem solving using simulations* (pp. 309-333). New York: Erlbaum.

von Davier, M. (2005). A class of models for cognitive diagnosis. *Research Report RR-05-17*. Princeton, NJ: ETS.

von Davier, A.A. & Halpin, P.F (2013, May). Modeling the Dynamics in Dyadic Interactions in Collaborative Problem Solving. Paper presented at an invited symposium at the annual meeting of the National Council of Measurement in Education, San Francisco, CA.

von Davier, A.A. & Halpin, P.F. (in press). *Collaborative problem solving and the assessment of cognitive skills: Psychometric considerations* [Research Report] Princeton, NJ: Educational Testing Service.

Vygotsky, L.S. (1978). Mind and society: The development of higher psychological processes. Cambridge, MA: Harvard University Press

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

Wainer, H., Dorans, N.J., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L., & Thissen, D. (2000). Computerized adaptive testing: A primer (second edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Walker, E., Rummel, N., & Koedinger, K. R. (2011). Designing automated adaptive support to improve student helping behaviors in a peer tutoring activity. *International Journal of Computer-Supported Collaborative Learning*, 6, 279-306.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (NISE Research Monograph No. 6). Madison: University of Wisconsin–Madison, National Institute for Science Education.

West, P., Wise Rutstein, D., Mislevy, R.J., Liu, J., Levy, R., DiCerbo, K.E., Crawford, A., Choi, Y., Chapple, K., Behrens, J.T. (2012). A Bayesian network approach to modeling learning progressions. In A.C. Alonzo & A.W. Gotwals (Eds.), *Learning progressions in science* (pp. 255–291). Rotterdam, The Netherlands: Sense Publishers.

Williamson, D., Mislevy, R.J., & Almond, R.G. (2000). Model criticism of Bayesian networks with latent variables. In C. Boutilier & M. Goldszmidt (Eds.), *Uncertainty in artificial intelligence 16*, pp. 634-643. San Francisco: Morgan Kaufmann.

Wilson, M.R. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin*, 105, 276-289.

Wilson, M.R. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, 46(6), 716-730.

Wilson, M.R. (2012). Responding to a challenge that learning progressions pose to measurement practice. In A. Alonzo & A. Gotwals (eds.), *Learning Progressions in Science* (pp. 317-343). Rotterdam: Sense Publishers.

Wolfe, E. W., & Gitomer, D. H. (2001). The influence of changes in assessment design on the psychometric quality of scores. *Applied Measurement in Education*, 14(1), 91-107.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. Brennan (ed.), Educational measurement (pp. 111-153). Portsmouth, NH: Praeger/Greenwood.

Zapata-Rivera, D. & Bauer, M. (2011) Exploring the Role of Games in Educational Assessment. In M.C. Mayrath, J. Clarke-Midura, & D. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 149-172). Charlotte, NC: Information Age.

About GlassLab

GlassLab brings together leaders in commercial games and experts in learning and assessment to leverage digital games as powerful, data-rich learning and formative assessment environments.

The Lab represents a groundbreaking collaboration between Institute of Play, the Entertainment Software Association, Electronic Arts, Educational Testing Service, Pearson's Center for Digital Data, Analytics & Adaptive Learning and others. With best-in-class talent and intellectual property from EA; trusted expertise in evidence-based assessment from ETS and Pearson; the ESA's distributed network of thought leaders and public advocates; and Institute of Play's expertise as a leading innovator in 21st century learning design,

GlassLab is creating a new model for commercial game studios and learning organizations to come together to do great work.

A project of Institute of Play, GlassLab is made possible through the generous support of The Bill and Melinda Gates Foundation and The John D. and Catherine T. MacArthur Foundation.

To learn more, visit www.glasslabgames.org
For the latest news, follow us on Twitter @GlassLabGames.