

Running head: EVIDENCE-CENTERED DESIGN OF EPISTEMIC GAMES

Evidence-centered Design of Epistemic Games:  
Measurement Principles for Complex Learning Environments

André A. Rupp, Matthew Gushta, & Robert J. Mislevy  
University of Maryland

David Williamson Shaffer  
University of Wisconsin at Madison

Address correspondence to:

André A. Rupp  
Department of Measurement, Statistics, and Evaluation (EDMS)  
University of Maryland  
1230 Benjamin Building  
College Park, MD 20742  
Tel: (301) 405 – 3623  
Fax: (301) 314 – 9245  
E-mail: [ruppandr@umd.edu](mailto:ruppandr@umd.edu)

## Abstract

We are currently at an exciting juncture in developing effective means for assessing so-called 21<sup>st</sup>-century skills in an innovative yet reliable fashion. One of these avenues leads through the world of *epistemic games* (Shaffer, 2006a), which are games designed to give learners the rich experience of professional practica within a discipline. They serve to develop domain-specific expertise based on principles of collaborative learning, distributed expertise, and complex problem-solving. In this paper, we describe a comprehensive research programme for investigating the methodological challenges that await rigorous inquiry within the epistemic games context. We specifically demonstrate how the *evidence-centered design framework* (Mislevy, Almond, & Steinberg, 2003) as well as current conceptualizations of reliability and validity theory can be used to structure the development of epistemic games as well as empirical research into their functioning. Using the epistemic game *Urban Science* (Bagley & Shaffer, 2009), we illustrate the numerous decisions that need to be made during game development and their implications for amassing qualitative and quantitative evidence about learners' developing expertise within epistemic games.

Evidence-centered Design of Epistemic Games:  
Measurement Principles for Complex Learning Environments

Learning in the 21<sup>st</sup> century is increasingly characterized by our ability to make and understand interconnections between concepts, ideas, and conventions across a variety of domains. Consequently, one of the principal challenges of our times is to adequately prepare learners of all ages for challenges in such an increasingly interconnected world, which is heavily permeated by the existence and use of digital tools. Various authors and institutions have proposed taxonomies of so-called *21<sup>st</sup>-century skills* that are believed to be at the core of the relevant expertise that is required for facing the demands of associated 21<sup>st</sup>-century tasks (e.g., Bagley & Shaffer, 2009; Partnership for 21<sup>st</sup> Century Skills, 2008; Shute, Dennen, Kim, Donmez, & Wang, in press). While there is no single definitive list of these skills, most lists focus on expanding traditional concepts of knowledge, skills, and abilities to encompass concepts such as critical and innovative thinking, systems-thinking, interpersonal communication and collaboration skills, digital networking and operation skills, intra- and intercultural awareness and identity, and cross-cultural sensibility.

Assessing 21<sup>st</sup> century skills frequently requires exposing learners to well-designed complex tasks, affording them the ability to interact with other learners and trained professionals, and providing them with appropriate diagnostic feedback that is seamlessly integrated into the learning experience. This can be accomplished within well-designed immersive virtual environments and related simulation-based learning environments. Their potential is increasingly realized by national funding organizations and private foundations, which are supporting concerted and rigorous research efforts into the effectiveness of these environments. For example, the *John D. and Katherine T. MacArthur Foundation* is funding various projects on game- and simulation-based learning through grants in their *Digital Media & Learning* initiative (<http://www.macfound.org>). At an international large-scale assessment level, the recently launched *Assessment and Teaching of 21<sup>st</sup> Century Skills Project* ([www.atc21s.org](http://www.atc21s.org)), co-sponsored by Cisco, Microsoft, and Intel, certainly represents the largest coordinated effort to date to develop worked examples for a variety of learning and assessment systems in this area.

It is critical to understand that the landscape of immersive virtual environments and related digital simulation tools is vast. Environments that might be specifically labelled as “educational

games” or “educational simulations” may be designed for a variety of purposes. As Clark, Nelson, Sengupta, and D’Angelo (2009) discuss, the collections of existing tools form their own genres whose primary purposes can vary from making learners aware of a particular issue, teaching them basic componential skills, or encouraging them to become professional scientists. The recently held *Learning Science: Computer Games, Simulations, and Education* workshop sponsored by the *National Academy of Sciences* ([http://www7.nationalacademies.org/bose/Gaming\\_Sims\\_Presentations.html](http://www7.nationalacademies.org/bose/Gaming_Sims_Presentations.html)) showcased the diversity in these environments and the research and program development surrounding them. The general focus of this paper is on a particular set of educational games, so-called *epistemic games*, and the lessons that can be learned from their design for creating meaningful assessment narratives about learners. Before describing the structure of our argument and organization of the paper in more detail, a few words about epistemic games and their design are in order.

### *Epistemic Games*

Discipline-specific learning as well as learning more generally is not simply restricted to the mastery of concepts and procedures, but includes the ability to think, act, and interact with others in productive ways to solve complex tasks in real-world situations. Becoming an architect, for example, is more than knowing materials properties and tools for computer-aided design. It is being able to see what architects see and being able to frame it in ways the profession thinks, knowing how to work with and talk with other architects and clients, and using concepts and procedures within the sphere of activities that constitutes architecture. In short, this is what is known as the *epistemic frame* of the discipline (Shaffer, 2006a, 2006b), which is what gives this genre of games its name.

Although there are many game- and simulation-based opportunities for transforming practices, perceptions, and commitments regarding learning in the 21<sup>st</sup> century (see, e.g., Gee, 2003; Gibson, Aldrich, & Prensky, 2006), epistemic games are explicitly based on theory of learning in the digital age. Specifically, epistemic games are digitally supported learning environments that are designed to allow learners to develop domain-specific expertise under realistic constraints (e.g., Bagley & Shaffer, 2009; Shaffer, 2006a). For example, learners may learn what it is like to think and act like journalists, artists, business managers, or engineers. This is accomplished by designing the game in such a way that completing it mimics the core experiences that learners outside the gaming environment would have in a *professional*

*practicum* in the field. The experiences that epistemic games afford and make accessible to learners are characterized by a blend of individual and collaborative work in both real-life and virtual settings.

Due to their focus on creating digitally supported learning experiences that adequately mimic the real-life complexities of a profession, epistemic games are different from other types of computer games in important ways. First, the development of expertise in a particular real-life profession is at the heart of playing an epistemic game while only elements of such expertise are typically developed as a side-product in commercial computer games that are designed for entertainment more generally. Notably, though, the objective of epistemic games is not to “train” learners with the purpose of suggesting particular career trajectories to them, but to facilitate the emergence of disciplinary thinking and acting that transfers to other contexts. Second, the decisions that are made in an epistemic game are made under real-life constraints and in real-time, which is contrary to computer games such as *SimCity* that allow the learner to manipulate time, resources, conditions, and decisions like an omnipotent being (Bagley & Shaffer, 2009; Shaffer, 2006a).

### *Designing Epistemic Games*

Clearly, designing an epistemic game is a highly complex task, requiring the reconciliation of instructional methods with principled assessment design, data collection, data analysis, score reporting, and formative feedback. As a result of these demands, an important tension arises: on the one hand, there is a need for high fidelity of the epistemic game vis-à-vis the real-life professional practicum that is traditionally used to train professionals in the field; on the other hand, there is a need for gathering reliable assessment data in support of the development of disciplinary expertise by the learners within the digital gaming environment. Core experiences of the professional practica must be offered to learners within the game, affording the opportunity to develop and demonstrate their epistemic frame, while concurrently providing the information necessary to satisfy evidence-based arguments regarding the development of said epistemic frame.

Put differently, while it may be possible to increase the fidelity of particular tasks with relative ease, the introduction of each new game element requires evidence extraction and accumulation rules that may require iterative fine-tuning that can be costly and resource-intensive. This is especially true if those game elements elicit complex and interdependent

behavior from learners. Similarly, while it may be possible to add additional assessment components to the game to gather supplementary data, this may disrupt the flow of the game play or present learners with tasks that feel unnaturally constrained with respect to the overall narrative frame that the game is using.

Two forms of data are collected from learners during the stream of epistemic game play in service of the above purposes, which we may label as *process data* and *product data* for the moment. Process data derive from interactions of learners with other learners as well as non-learners (i.e., instructors / mentors) while product data derive from the collection of the learners' tangible work products. This rich, dual-mode data stream presents obstacles for traditional assessments. For example, learner activity is highly contextualized and, therefore, observations are unlikely to be independent. Furthermore, the assessment design as the environment is less constrained than traditional assessment environments, which allows for more unexpected student responses.

These challenges are akin to those encountered from complex performance-based assessments (see, e.g., Williamson, Mislevy, & Bejar, 2006) except that the data collected within electronic game environments is even more multilayered than within non-electronic environments. The game design needs to ensure that the tasks and opportunities for interaction are of sufficient richness and flexibility to allow learners to engage their epistemic frames just as trained professionals would. Therefore, the game design needs to ensure that the data can be collected relatively unobtrusively so as to not change the task demands. To address these demands, some researchers in the field of game-based assessment have argued for leveraging unobtrusive data-collection efforts in so-called *stealth assessments* (Shute & Spector, 2008; Shute, Ventura, Bauer, & Zapata-Rivera, 2008).

Clearly, then, designing epistemic games that meaningfully capture learner performance requires bringing together a series of experts in the targeted domain, digital information technology, performance-based assessment, and multidimensional statistical modelling. As with any well-designed assessment, a principled approach to game design based on a framework that can accommodate both the fidelity and the assessment demands is a methodological necessity. A desirable framework ensures that the process of game design is well-structured, decisions are clearly articulated, and linkages between individual game components are explicitly tied to the desired narratives about learners' developing expertise.

The educational measurement literature provides ample descriptions of key principles of assessment design for more traditional knowledge-based assessments such as mixed-format achievement tests and more traditional performance-based assessments such as writing tests and portfolio evaluations (see, e.g., Downing & Haladyna, 2006). More recently, systemic quality control processes for managing assessment systems have been described (Wild & Ramaswamy, 2007), as well as principled frameworks for designing and evaluating performance in simulation environments (Baker, Dickieson, Wulfbeck, & O'Neil, 2007). Despite the richness of this knowledge base, much of the information remains either marginally relevant or too compartmentalized for game-based assessments. In short, designing epistemic games with an assessment lens requires careful adaptation of existing design principles for complex performance-based assessments.

### *Evidence-centered Design*

To support assessment developers in making explicit the rationales, choices, and consequences reflected in their assessment design, the framework of *evidence-centered design* (ECD) was created (for an overview see, e.g., Mislevy, Almond, & Steinberg, 2003; Mislevy, Almond, Steinberg, & Lukas, 2006). While ECD can, technically, be applied to the development of any kind of assessment where the a priori definition of constructs and associated variables is meaningful, it is particularly suitable to the development of performance-based assessments that are created in the absence of easily delineable test specifications. It is in these contexts that the number, complexity, and connectedness of decisions that need to be made about the assessment design are most daunting. Moreover, because developing such assessments is costly and time-consuming, these are also the contexts in which there is a strong need for *re-usable design templates* whose structure can be linked systematically to particular sets of statements that are made about learners (e.g., Serataan & Mislevy, 2009; Brecht, Cheng, Mislevy, Haertel, & Haynie, 2009; see Plass, Homer, & Hayward, 2009). Because it provides a systematic way for addressing these desiderata, ECD is a natural bridge between the two demands for high fidelity and rich assessment data that epistemic games have to address.

### *Purpose of this Paper*

In this paper, we lay out key assessment principles for designing epistemic games within an ECD framework. In doing so, we illustrate how ECD can serve to structure and connect the decisions that are made at various points during the development of an epistemic game. Since

research in this area is in its infancy, relatively speaking, the objective of this paper is not to provide a set of definitive answers. Rather, it is to sketch out a methodological research programme for investigating how epistemic game design influences the data that are gathered from epistemic games and how these data can be analyzed to appropriately capture the developing expertise of different learners. Our research teams at the University of Maryland and the University of Wisconsin at Madison are currently engaged in empirically addressing a variety of the research avenues that we chart in this paper through several NSF-funded research grants. We have written this paper to inspire others to ask systematic questions about epistemic game research and to engage in this exciting new field of research.

Rather than merely discussing the main components of the ECD framework in an abstract manner, we use an actual epistemic game to illustrate these key principles and the associated methodological challenges. The epistemic game that will be used for illustration purpose is *Urban Science*, which mimics the professional practicum experiences of urban planners and is described in the Appendix. The game is developed at the University of Wisconsin at Madison (see <http://epistemicgames.org/eg/> for more information), which is continually updated based on empirical data from local implementations. It is also used as the leveraging environment for the NSF-funded *AutoMentor* and *Dynamic STEM Assessment* grants, whose goals are to develop automated feedback mechanisms for epistemic games and to research the utility of different measurement approaches for these gaming environments, respectively.

We have divided this paper into three main sections. In the first section, we describe basic assessment concepts and resulting design principles that are relevant for the development of performance-based assessments in general and epistemic games in particular. These include the key concepts of reliability / measurement error as well as validity / validation. Our description culminates in a more detailed presentation of the ECD framework, whose implementation addresses these concepts from a different angle and, arguably, unifies an evidentiary argument for them in a practical manner. In the second section, we describe the kinds of decisions that need to be made within each of the ECD model components within an epistemic game context and illustrate these decisions within the real-life context of *Urban Science*. Using ECD effectively addresses validation research from a procedural development perspective. In the third section, we articulate how additional validation research for epistemic games can be structured to address the key validity aspects presented in the second section of the paper. Using this approach

effectively addresses validation research from a more compartmentalized perspective, which can be meaningfully overlaid with the ECD perspective. We close the paper with a look at the implications of the argument that is presented in the paper.

### Principled Assessment Design under an ECD Lens

In this first main section of the paper we review two key assessment concepts and the resulting design principles that are necessary for understanding the discussions that follow in the two subsequent sections. The two concepts are *reliability / measurement error* and *validity / validation*, whose basic premises apply to any kind of performance assessment generally and epistemic games in particular. We specifically show how addressing these characteristics for epistemic games is rather complex and quickly transcends their seemingly straightforward definitions for traditional forms of assessments in the educational measurement literature.

Notably, the bulk of the statistical machinery for studying reliability and validity has been developed under the psychological perspectives of trait and behavioural psychology. This does not limit the relevance of these two concepts to assessment systems that are designed under such perspectives, however. Quite to the contrary, the two concepts challenge us to answer fundamental questions of evidence that are broadly applicable to all types of assessments: How can we gather trustworthy evidence in unfamiliar contexts such as games, collaborative work, and interactive environments? Where can we adapt tools from educational and psychological measurement to these new settings? How can we use these adapted tools to help us evaluate and improve the efficiency, feasibility, and effectiveness of our learning environments (Moss, 1994; Mislevy, 2004)? In short, they help us to frame fundamental questions about the quality of the assessment data and the resulting interpretations that are relevant to all statements that are made about learners.

#### *Reliability and Measurement Error*

*Reliability*, which is broadly defined as the *consistency* of some event, is a plausible concept that people use frequently in everyday life (see Mislevy & Braun, 2003). In an assessment context, reliability refers specifically to the consistency, across assessment conditions, of the score patterns that are used to say what learners know and can do (e.g., Frisbie, 1988). The key requirement for consistency is thus *replication* because only through replications of events can we talk about the consistency of observed patterns across these events (Brennan, 2001a). In epistemic games, the concept of reliability challenges us to ask

how consistent the scores for learners' performances are under different gaming conditions (e.g., different tasks, different sequencing of tasks, different mix of tasks, different platform designs), and at what level to aggregate patterns of variation with a particular statistic.

*Measurement error* is a quantification of the amount of uncertainty of unobservable learner characteristics that is associated with the observed scores that we see. In the context of epistemic games, the unobservable characteristics of learners would be the sets of 21<sup>st</sup> century skills that the games are targeting. We attribute this inferential uncertainty to the fact that the epistemic games, viewed as assessments, are fallible instruments and can be used only imperfectly to make statements about learners. Conceptually, measurement error is inversely related to reliability such that statistics that are more reliable have less measurement error and vice versa.

Despite these relatively intuitive definitions of reliability and measurement error, defining and estimating reliability and measurement error in statistical models can become complex rather quickly (see, e.g., Allen & Yen, 2002; Brennan, 2001a). In part, this is due to the fact that unobservable characteristics are traditionally represented by *latent variables* in statistical models, which, in contrast to observed variables, account for measurement error in the assessment procedure. For example, Brennan (2001b) discusses how one can operationalize, quantify, and decompose measurement error from the perspective of *classical test theory (CTT)* and *generalizability theory (g-theory)* while Embretson & Reise (2000) and de Ayala (2009) do the same from an *item response theory (IRT)* perspective.

Formalizing reliability and measurement error within a single modelling framework such as CTT, g-theory, or IRT can sometimes be complicated and subtle, but translating the concept of reliability across different measurement frameworks can prove even more challenging (e.g., Adams, 2006; Templin & Henson, 2009). However, defining and quantifying score reliability and measurement error is absolutely necessary to understand whether the interpretations that are made about learners' are trustworthy. Within the context of epistemic games, this may require novel definitions, studies, and reporting strategies for reliability and measurement error. In the end, the resulting statistics may even turn out to be game-specific just as is the case with scoring systems for other complex assessments (e.g., Williamson, Mislevy, & Bejar, 2006).

Defining and quantifying reliability and measurement error of scores is particularly challenging in the context of epistemic games for three principal reasons. First, it is very challenging to minimize measurement error from an assessment design perspective due to the complexity of the tasks, which induce dependencies across observations. Second, there are multiple layers of human judgment involved in generating observed indicators whether the scoring is automated or not. Third, given the relative richness of the tasks and the resulting observations, process and product data are collected at a level of detail that may be relatively distal to the desired interpretation of the responses.

### *Validity and Validation*

Validity is the traditional complement to reliability. Modern conceptions view *validity* as a property of the interpretations that are made about learners, rather than as a property of the scores themselves. Consequently, validity has to be assessed with reference to each particular interpretation that is desired about learners, which requires that these interpretations are clearly described and organized before an assessment is developed. In traditional assessments, interpretations are often cast in terms of latent trait or behavioural perspectives but this need not be the case. For instance, socio-cultural or socio-cognitive perspectives can be equally defensible (Mislevy, 2008), which are more appropriate for epistemic game contexts.

It is generally agreed upon that validity is not an absolute but, rather, a matter of degree such that interpretations can be ascribed a particular *degree of validity* at a particular point in time. As prominent researchers such as Kane (2007, 2008) remind us, however, thinking about validity does not require thinking about all potential interpretations that could ever be made by anyone on the basis of the entirety of assessment data. Rather, it requires thinking about the interpretations that are made for the particular *purpose* the assessment is designed for or the particular *use* to which it is put in alignment with the purpose.

Modern conceptions view validity as a *process* that is on-going. At the same time, most practitioners would probably argue that there are moments in time when a certain level of *evidentiary saturation* is reached that does not seem to require further collection of evidence to support particular interpretations at that point. Clearly, however, validity is a function of the *evidentiary frameworks* of the stakeholders who utilize and defend the interpretations as well as the *disciplinary standards* within which the interpretations are framed.

Despite the fact that inferential validation is seen as an evidence-based endeavour that is aimed at unifying the evaluation of interpretive narratives, most practitioners find it helpful to distinguish different facets or aspects of validation. For example, Messick (1995) lists the following seven facets of validity or validation processes (see also Messick, 1989), which can guide our investigations of validity of interpretations in the context of epistemic games:

1. Content validity  
⇒ does the content of the assessment represent the target domain?
2. Substantive validity  
⇒ do the respondents engage in the appropriate cognitive processes?
3. Structural validity  
⇒ does the scoring process reflect the interaction of abilities in the domain?
4. Predictive validity  
⇒ can the assessment scores be used to predict an outcome of interest?
5. External validity  
⇒ do respondents perform similarly on assessments tapping similar constructs and differently on assessments tapping different constructs?
6. Generalizability  
⇒ can the assessment results be generalized across different conditions such as time points, administration contexts, and respondent samples?
7. Consequential validity  
⇒ do the assessment interpretations lead to fair and defensible consequences for respondents?

While the view detailed above is the predominant view on validity / validation, some researchers dislike any subjective perspectives on validity that rely on human judgment and evaluation. They argue instead that validity should be cast as an objective measurable property of assessments that is tied to the notion of causality (Borsboom & Mellenbergh, 2007). Given that epistemic games rely heavily on socio-cognitive and socio-cultural theories of learning rather than purely information-processing perspectives from cognitive psychology, however, such a restricted causal perspective is probably too narrow to be of practical use for research on epistemic games.

As we shall see below, what makes validation in the context of epistemic games rather challenging is the complexity of the desired interpretations, because statements about learners' expertise is framed longitudinally and depends heavily on the evolving interrelationships between their core latent characteristics. Addressing these challenges is quintessential to making statements about learners' developing expertise defensible to a wider audience, however,

especially if the audience has a critical assessment eye. It is relevant to note at this juncture that there are currently no disciplinary standards that we are aware of for what the exact nature of evidence for reliability and validity would have to look like in the context of epistemic games. Nevertheless, the *Standards for Educational and Psychological Testing* (APA, NCME, & AERA, 1999) and the related *Code of Fair Testing Practices* (Joint Committee on Testing Practices, 2004) provide guidelines for thinking about the quality of evidence in an epistemic game context. For example, questions about the fairness of tasks for different subgroups of learners, effective feedback mechanisms that appropriately scaffold learning progressions, and the impact of the interface design on the nature and quality of learner performances are quintessential to ask.

#### *Models in the ECD Framework*

The discussions about reliability / measurement error and validity / validation in the previous section have shown that the demand for evidence on the trustworthiness of score profiles and the resulting defensibility of the interpretations about learners in assessment is generally high. The discussions have hinted at the high degree of interrelationship that exists between a myriad of decisions that need to be made during the design, implementation, scoring, and reporting process for an assessment to address evidentiary demands by stakeholders.

The ECD framework was developed to help assessment designers structure their thinking and their actions to address this complex task. As a preface, the entire assessment design process is driven by the process of *domain modeling*. This is particularly important for principled assessment design, because an assessment argument is laid out in narrative terms at this stage: Just what is it that we would like to say about students? What kinds of things do we need to see them say, do, or make, in what kinds of situations? In more formal terms, based on Toulmin's general structure for argument, Mislevy (2006) argues that these questions form the keel of assessment design no matter what kind of evidence or interpretations are required, or what psychological perspective the assessment is used to frame the argument.

The ECD framework then identifies different layers and models at which different kinds of activities take place, as shown in Figure 1. These activities include modeling the target domain via appropriate tasks, assembling the tasks into a coherent assessment, and delivering the assessment with suitable interfaces. They specifically consist of (1) the *student models*, (2) the *task models*, and (3) the *evidence models*, which form what is known as the *conceptual*

*assessment framework*. These models are then glued together by (4) the *assembly model* and (5) the *presentation model* that make the assessment deliverable.

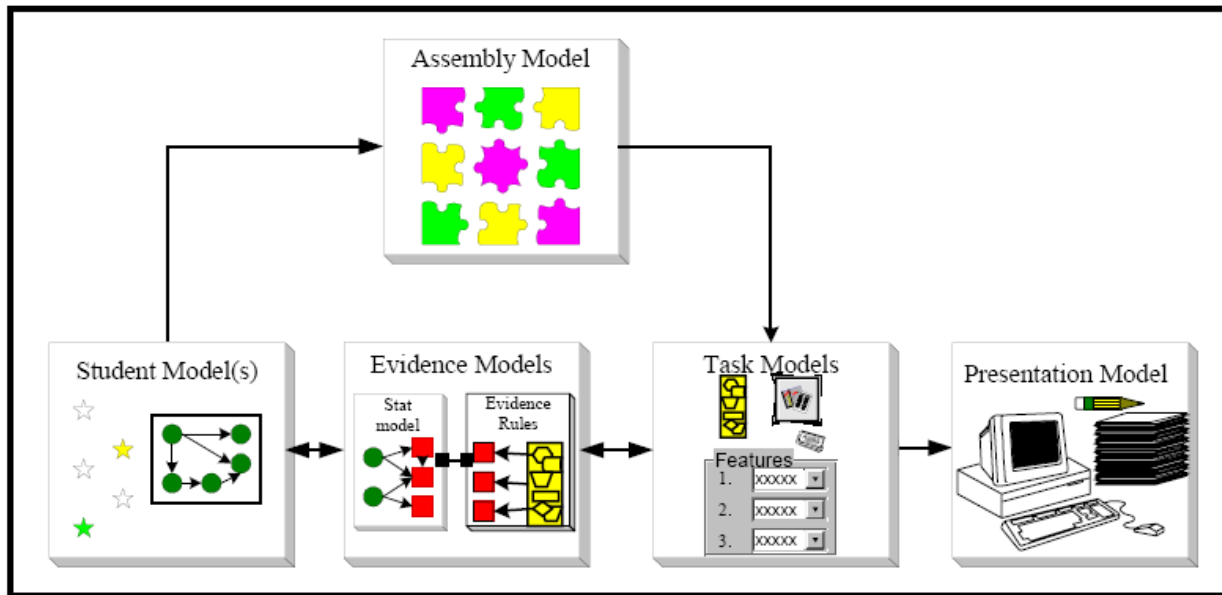


Figure 1. A schematic representation of the models in the ECD framework (Mislevy et al., 2006).

The *student model* provides a proxy model for the proficiency structure(s) of learners as motivated by learning theory. It is the elements of that proficiency structure to which relevant observed variables are then “docked” via statistical models and it is domain-specific theories about mental models that make such linkages defensible. In the context of epistemic games, theories about the key elements of the epistemic frame, the way in which they are interconnected, the way in which they are engaged in particular tasks, and the way in which their interconnections change over time feed into the construction of the student models.

The *task models* then specify the conditions and forms under which data are collected. Variables in a task model are motivated by the nature of the interpretations the assessment is meant to support and may include information about the context, the learner’s actions, and the learner’s past history or particular relation to the setting. All of the variables may be required to make sense of learners’ actions in the situation in which they were made.

The *evaluation component* or *evidence rules component* of the *evidence models* specifies the salient features of whatever the learner says, does, or creates in the task situation, as well as the rules for scoring, rating, or otherwise categorizing the salient features of the assessment. The

*probability component* or *statistical model component* of the evidence models specifies the rules by which the evidence collected in the evaluation is used to make assertions about the student model; in other words, it is the place where the measurement model for particular tasks is specified.

The *assembly model* describes how these different components are combined for answering particular questions about learning in a given assessment situation. Finally, the *presentation model* describes whether modes of task and product presentation change across different parts of the assessment and what the expected implications of these changes are. In practice, ECD models for a given assessment are constructed jointly and refined iteratively, because the full meaning of any model in the framework only emerges from its interrelationship with other components.

Committing to an ECD framework is, thus, a particular way of committing to *principled assessment design*. It translates the theoretical treaties on reliability and validity along with certain rudimentary guidelines for the development of performance assessments (e.g., Bennett Jr., Lance, & Woehr, 2006; Hale, 1999) in a systematic way. Since epistemic games can be viewed as a particular type of complex performance assessment, it can be beneficial to apply the ECD framework to the design, implementation, and reporting of epistemic games. The ECD framework can help to make explicit how the demands for high fidelity and rich assessment data in epistemic game contexts are addressed through a myriad of design, implementation, analysis, and reporting decisions. It can also underscore what the implications are of these decisions for statistical measurement. Finally, it can provide a roadmap for thinking about research programs that investigate how traditional measures of reliability / measurement error and validity / validation can be adopted, adapted, and extended for epistemic games. These theoretical benefits of ECD were the key motivation for why we chose to adopt this framework for structuring our research programme for epistemic games. In the next section of the paper, we use the ECD framework to describe several key decisions that need to be made during the design, implementation, and analysis of epistemic games.

#### Decisions across ECD Models in Epistemic Games

For the purposes of this paper, we illustrate our thinking with research and development surrounding *Urban Science*. That is, we describe the kinds of decisions that need to be made within each model of the ECD framework and illustrate some representative decisions made for *Urban Science*. Whenever possible, we also address the interrelationships between decisions

made in different models. Our goal for the remaining sections is to describe a set of research questions rather than a set of comprehensive empirical answers, focusing on the work that is currently conducted through our research programme. Recall from the previous section that the core ECD models that reflect key domain modelling decisions are (1) the student models, (2) the task models, (3) the evidence models, (4) the assembly model, and (5) the presentation model, which we are going to discuss in this order.

### *Student Model Decisions*

The development of an epistemic game is driven by a theory about learning within the domain that describes emerging levels of expertise that are reflected in distinct patterns of acting and thinking as professionals that are driven by their epistemic frame. These theories are grounded in what can be called the *epistemic frame hypothesis* (Shaffer, 2006b), which suggests that any community of practice has a culture whose base structure is composed of:

1. *Skills*: the things that people within the community do
2. *Knowledge*: the understandings that people in the community share
3. *Identity*: the way that members of the community see themselves
4. *Values*: the beliefs that members of the community hold
5. *Epistemology*: the warrants that justify actions or claims as legitimate within the community

This collection of *skills, knowledge, identity, values, and epistemology (SKIVE)* forms the epistemic frame of the community. The epistemic frame hypothesis claims that: (a) an epistemic frame binds together the skills, knowledge, values, identity, and epistemology that one takes on as a member of a community of practice; (b) such a frame is internalized through the training and induction processes by which an individual becomes a member of a community; and, (c) once internalized, the epistemic frame of a community is used when an individual approaches a situation from the point of view (or in the role) of a member of a community.

The epistemic frame for *Urban Science* consists of domain-specific frame elements that are organized into the five SKIVE categories. These categories were derived from Ehrlich (2000) and the *National Assessment Governing Board's* (2006) civics framework:

1. *Skills (various)*: being able to communicate clearly, both orally and in writing; being able to collect, organize, and analyze information; being able to think critically and justify different positions; being able to view issues from the perspective of others.
2. *Knowledge (terms of art, systems thinking)*: knowing institutions and processes that drive civic, political and economic decisions; knowing how a community operates, the problems it faces, and the richness of diversity.
3. *Identity (as planner, as professional)*: having a way of seeing oneself that is commensurate with how members of the urban planning community see themselves.
4. *Values (working for stakeholders, for the public good, as a team, like a professional)*: being willing to listen to, and take seriously, the ideas of others.
5. *Epistemology (general, planning-specific)*: being able to understand what counts as relevant evidence that justifies actions as legitimate within the urban planning community.

All of these frame elements are interconnected and, taken together, form the epistemic frame of urban planners. Thus, the epistemic frame hypothesis predicts that the elements and their interrelations become internalized through appropriate training and immersion in the urban planning profession. The structure of these experiences needs to encourage learners to activate task-appropriate frame elements that guide how they think and act in urban planning contexts, a process that is mimicked through *Urban Science*.

Put in more concrete terms within the context of urban planning, professionals in the domain act and reason like urban planners, identify themselves as urban planners, are interested in urban planning, and know about geography, architecture, mathematics, information technology, and other relevant technical fields. The same is true for other professionals such as architects, policy-makers, city council members, journalists, business managers, of course, but reflected in different ways of thinking anchored in different epistemic frames.

From a student model perspective, decisions need to be made about the number of latent characteristics that should be modelled for each learner, which are the components of the epistemic frame that need to be statistically represented in a model. But how many such skills should be explicitly modelled in an analysis of learners' performance? The epistemic frame for urban planners as described above suggests that there should be at least five latent characteristics (i.e., skills, knowledge, values, identity, epistemology), as well perhaps as some characterization(s) of their interconnectivity. As these labels show, however, any interpretations about learners using this level of granularity will be rather coarse, because a variety of sub-skills, types of knowledge, kinds of values, aspects of identity, and characteristics of epistemology are

subsumed under these labels. Hence, it may be desirable to model these frame elements at finer levels of grain-size as indicated in the list above. To decide on a particular representational level of the epistemic frame, a theory about learning in the profession is required that is evaluated vis-à-vis the decisions that need to be made about learners on the basis of the game data.

In some sense, this boils down to deciding whether a finer differentiation of the core frame elements is a matter of quantitative differences or qualitative differences for the purpose at hand. For example, it seems plausible to view the skills of “being able to communicate clearly in writing” and “being able to view issues from the perspectives of others” as qualitatively distinct, albeit related, skills. At the same time both of these skills can be mastered to different degrees, which implies that modelling them could be done using either continuous proficiency indicators or discrete proficiency indicators.

However, as we discuss below in the evidence model section, the degree to which statistical models can model a large number of frame elements in a reliable manner depends on the amount of unique information that is available about each element through the game and the complexity of the statistical model under consideration. Therefore, certain theoretically or practically desirable levels of differentiation may be impossible to represent reliably with particular statistical models. As with other forms of assessment, the level of representation that is used for reporting purposes is typically based on a statistical model that is parsimonious, which means that it contains the least complex structure that is necessary to capture the essential patterns in the data so that it can sufficiently support desired evidentiary narratives and decisions.

On the surface, it may seem that SKIVE elements are similar to latent characteristics measured in more familiar assessments. For example, professional knowledge in general may be characterized by the use of domain-appropriate terms of art or specialized language for describing situations and actions. In cases such as these, existing techniques from simulation-based assessment might be adapted for use in the assessment of epistemic games, which has been part of previous research into the effects of epistemic games as learning tools (e.g., Bagley & Shaffer, 2009).

But assessing the development of an epistemic frame is more complex because such a frame implies not only the presence of such elements, but a particular interconnectivity among these SKIVE elements as students bring them to bear in the evolving situations of the game in which they are relevant. Both interconnectivity and its changing nature throughout the course of the

game are more novel aspects of assessment in the context of epistemic games. Phrased differently, it is the longitudinal focus on changes of association structures over the course of the game play in addition to graded mastery statuses in the context of a complex performance-based assessment that make the assessment of SKIVE elements in epistemic games particularly challenging.

Finally, since a frame of reference is required to evaluate any given state of the epistemic frame of a player at any point in time throughout the game, such decisions would have to be made in light of describing novices and experts with respect to the multiple frame elements and the ways they are expected to act when confronted with particular tasks. Which combinations of mastery levels for different skills are legitimate representations of novice and expert urban planners that can be used to structure characterizations of learners is an important question. Answering this question requires rational evidence from theory and empirical evidence from qualitatively grounded ethnographic studies. The field is just beginning to amass such evidence (e.g., Bagley & Shaffer, 2009).

#### *Task Model Decisions*

The tension between the need for a high fidelity experience within *Urban Science* vis-à-vis the real-life practicum experiences it mimics as well as the richness of assessment data it generates is addressed according to decisions made in the context of task models. From a principled assessment design perspective, it is useful to encode *task design variables* for each task in an epistemic game. Task variables describe the key characteristics of the situations within which the learners act and the objectives of the tasks. Important task model variables include, for example, the targeted level of behaviour, the complexity of the task, potential strategies suitable for the task, and the level of support a learner has been provided for the use of SKIVE elements. Given that the opportunities for relatively unobtrusive data-collection are richer than in non-gaming contexts, appropriate task variables that predict performance can be used in statistical models to create more reliable scores that help to adequately predict task performance (see de Boeck & Wilson, 2004).

#### *Product versus process data revisited*

In *Urban Science* core product data consists of (a) entries into a planning notebook, (b) preference surveys, (c) final redevelopment plans, (d) final reports, and (e) final presentations of the reports. Each of these products includes verbal as well as visual information. Process data

can generically be viewed as any data that is recorded during the process of solving a task that is not at all, or only partially, captured in the product itself. In this sense, entries into a planning notebook are partially product and process data.

Interactional process data specifically consist of transcripts of (i) interactions between different learners and (ii) interactions between learners and mentors. The latter interactions consist of four sub-types characterized by who is involved in the interactions and whether they are planned by design or not. Thus, there are (i-1) interactions between mentors and individual learners at pre-determined points in the game, (i-2) interactions between mentors and groups of learners at pre-determined points in the game, (i-3) interactions between mentors and individual learners at learner-initiated moments in the game, and (i-4) interactions between mentors and groups of learners at learner-initiated moments in the game.

The use of mentors is an essential component of epistemic games, because mentors are key in scaffolding learners' experiences in professional practica. However, the manner in which mentors engage with learners is critical because the way in which they probe learners about their thoughts, actions, and rationales influences the quality of the observable data that are available. We are currently investigating how the strategies that mentors use in interacting with learners influences the nature of the resulting data and how different feedback mechanisms could be tailored to learners at different levels of developing expertise. For example, novice learners in epistemic games seem to require more frequent, more targeted, and less elaborate mentor feedback while expert learners seem to require less frequent, broader, and more elaborated feedback (see Shute, 2008).

Importantly, tasks have to be designed that elicit sufficient information about the SKIVE elements at the desired level of granularity. The nature of tasks has implications for the strength of evidence that resulting data provide about learners' developing expertise. For example, it is relatively straightforward to ask learners to collect and analyze data from different sources that are provided to them and code the resulting products using expert-generated rubrics. That is, in line with familiar assessment practices, product data can provide some evidence about the activation of frame elements, especially if intermediate drafts of work products are analyzed in addition to the final work products. Little evidence generally remains, however, of the on-line knowledge and skills employed during the activity that could be used to gather evidence about the nature and depth of interconnections among the frame elements.

Much useful information for learning about epistemic frame elements is thus potentially contained in the process data, rather than the product data. Since many, but not all, tasks in epistemic games are collaborative, however, the degree and nature of scaffolding is critical as it influences the interactional structure and, therefore, the contributions of individual learners toward task completion. That is, when learners act collaboratively with others to solve tasks in epistemic games, decisions need to be made about how to separately trace the intellectual and practical contributions of individual learners and the group as a whole.

*Coding of process data and automated feedback*

Furthermore, it is empirically rather challenging to collect data on a SKIVE element such as whether learners “take the ideas of others seriously.” As in other performance assessment contexts, capturing traces is much easier than judging which of the statements that learners make in these traces constitute reliable evidence for an activation of the desired frame elements. Delineating such evidence requires detailed coding schemes for tasks and resulting traces that take into account (a) the framing of the questions that led to the traces, (b) the rhetorical structure of each trace so that coding redundancies are avoided, and (c) the tendencies of individual learners to produce detailed traces.

That is, the interpretation of learners’ actions heavily depends on context in epistemic games. From a statistical perspective, tools from the areas of document and discourse analysis, data-mining, and natural language processing, coupled with suitable multivariate data analysis tools, can be leveraged meaningfully to aid in the interpretation of process data. For example, in collaboration with Art Graesser and other researchers at the *University of Memphis* we are currently exploring the utility of latent semantic analysis (Landauer, McNamara, Dennis, & Kintsch, 2007) for identifying meaningful clusters of utterance components that can be used as the basis for automated coding routines that can eventually replace mentor-generated feedback for learners during the game (see Storey, Kopp, Wiemer, Chipman, & Graesser, in press).

*Segmenting process data*

While product data are physically bounded, process data are open to decisions about *segmentation*. For example, one can define objective segmentation boundaries based on characteristics that can be measured without error such as blocks of time (e.g., every 15 minutes) or interactional boundaries (i.e., the beginning and end of a conversation). However, one can also define subjective segmentation boundaries based on consensually derived characteristics such as

thematic units (e.g., every conversation segment focused on a preference survey) and procedural units (e.g., every conversation segment that focuses on constructing something, rather than reflecting on something) whose coding process is prone to error.

How segmentation boundaries are defined is critical for subsequent analyses, because different statistical measures that are defined at the segment level under different statistical models are differentially sensitive to segmentation boundary shifts that arise from alternative coding approaches. Initial research in this area has shown that these segmentation boundaries are important and that different classes of statistics are differentially sensitive to this segmentation. For example, the absolute value of certain statistics for individual SKIVE elements are more strongly affected than the relative weights of the SKIVE elements, which appears to be reasonably robust (e.g., Rupp et al., 2009b).

#### *Temporal disjuncture of process data*

An important second-order issue with which the researchers have to contend is the temporal disjuncture between action and reflection. In the course of recording process or product data, for example, learners may refer to an action that took place at a time previous to the recording of the data: “Well, before I came up with this version, I tried to increase green space by removing lots of the parking, and I liked it, but most of the stakeholders didn’t.” Such a statement could be taken to make assertions about two points in time: the time at which the statement was made, and the previous time at which the reported event occurred. Segmentation in such a circumstance is more complex than merely delineating boundaries between events. In the case of epistemic games, researchers create a *play history* that associates pieces of data with their time-relevant referent or referents, and it may be this second-order artefact that is segmented, rather than the original data itself.

#### *Informational ambiguity of process data*

Some linguistic segments may also contain evidence about multiple frame elements at the same time so that the elicitation of such statements in interactions becomes particularly critical. For example, a statement such as “I know that I need to bring together the messages from the different environmental groups for the council member” shows as much evidence of a skill (i.e., developing a report) as evidence of values (i.e., taking into account the perspective of others). In other words, process data are challenging to collect and analyze, because the richness of the data depends on the opportunities that are given to each learner to express them, the defensibility of

coding categories applied to the data, and the reliability of the coding, if done by completely or partially by human raters.

### *Evidence Model Decisions*

ECD underscores, but does not overemphasize, the importance of the statistical models that are used in the evidence model component. A statistical model is viewed as the grammatical architecture of the resulting narrative about learners' developing epistemic frame and has a serving role in the creation of this narrative. While it is at the core of the creation of this narrative, it is not the focal point for the assessment design.

#### *The number of latent variables*

The decision about the number of latent characteristics that should be modelled from the student model has direct implications for the statistical model that is selected. If only interpretations about the five primary latent characteristics in the SKIVE framework are desired, for example, the statistical models need to include only five latent variables, no matter what statistical model is chosen. If the objective is to differentiate between these frame elements at a finer level of grain-size, however, then additional latent variables need to be included in the model. For example, if the objective is to distinguish only two components for each frame element, the number of latent variables that are required already increases from five to ten.

While this may not seem problematic generically, the literature on multidimensional latent variable models has repeatedly shown that it is numerically very challenging to differentiate between multiple latent dimensions (e.g., Rupp, 2008). Most latent variable models that are estimated successfully with real data sets contain between four and six latent variables at the most. Moreover, when the latent variables represent multiple latent characteristics within a single domain, they are often highly correlated and may also produce relatively unreliable scores (e.g., Haberman, 2008; Sinharay, Haberman, & Puhan, 2007). In such cases, little unique information is provided by each latent variable and any resulting profiles for learners, even though they can be numerically computed, they may be statistically rather unstable.

#### *Modeling context effects*

Again, it is also important to consider that responses of individual learners are highly context dependent in epistemic games because learners solve complex tasks that contain various interrelated elements and common stimuli. For example, at some point in *Urban Science* teams of learners representing multiple stakeholder groups have to work together to find an optimal

zoning solution for the neighbourhood of interest. This requires learners to investigate the impact of zoning proposals on environmental indicators, which are iteratively updated and displayed in an interactive simulation interface that supports this decision-making process. Traditional latent variable models are rather sensitive to such dependencies and a lot of research effort has gone into including statistical accommodations for such effects. For example, the literature provides for testlet models in the area of item response theory (e.g., Wainer, Bradlow, & Wang, 2007), Bayesian inference networks (e.g., Almond, Mulder, Hemat, & Yan, 2006), or diagnostic classification models (e.g., Rupp, Templin, & Henson, in press). Statistically speaking, since uncertainty about the epistemic frame of a given learner at a given point in time is partially due to the dependencies in task material and partially due to the dependencies arising from collaborative efforts. Those sources of variation in the scores should, ideally, be statistically disentangled while multiple reliable dimensions are created.

*Modeling the longitudinal data structure parametrically*

Epistemic games also pose an important additional challenge for the statistical modelling of the data due to their developmental focus within a relatively short time span. From a traditional latent variable perspective, interrelationships are operationalized as correlations between latent variables. Yet, assessment data are typically thought of as collected at one point in time in their entirety, rather than successively over time in a piecemeal fashion. Data that are collected longitudinally are typically used to summarize score profiles for groups of learners, rather than individual learners. Statistical models that model correlation structures with latent variables such as growth mixture models or curves-of-factor models (e.g., Duncan, Duncan, & Strycker, 2006; Fitzmaurice, Davidian, Verbeke, & Molenberghs, 2008) make rather strong assumptions and frequently require relatively large sample sizes.

Furthermore, most models for longitudinal data analysis are designed for inter-individual comparisons and may not necessarily represent the intra-individual latent-variable structure accurately (Bauer, 2007). However, it is the latter that is of most interest for the analysis of epistemic game data. Thus, while statistical models for longitudinal data can provide trustworthy inferential statistics if they can be estimated well, they are more theoretically appealing than practically useful at this point in time when epistemic games are implemented at comparatively small scales (Rupp et al., 2009a).

*Modeling the longitudinal data structure non-parametrically*

Several different avenues have been pursued to address these challenges with traditional latent-variable models. First and foremost, perhaps, is the use of non-parametric statistical methods that weaken the assumptions of traditional latent-variable models and are based on algorithmic methods rather than full-fledged integrated estimation approaches. Because algorithmic methods do not make assumptions about data that are as strong as statistical models embedded in a fully probabilistic framework, they can support a higher-dimensional representation of latent characteristics. This comes at a price, however, which is that inferential statistics that help to bound uncertainty of statements about these characteristics are often not available.

Currently, data from Urban Science are modelled using methods from *epistemic network analysis (ENA)* (Shaffer et al., 2009), which is rooted in methods from non-parametric statistics and social network analysis. The codes for the process and product data that were described in the task model section above are summarized first by creating *adjacency matrices*. These matrices are well-known in educational measurement (e.g., Tatsuoka, 1983) and indicate, via binary indicators, which skills co-occur with one another.

In *Urban Science*, the adjacency matrices are statistically generated from the indicator codes for the SKIVE elements at each time slice. For example, if the code sequence is an adjacency list of the form [1, 1, 0, 1, 1], indicating that a learner used four out of the five SKIVE elements in a particular time slice, then the adjacency matrix would contain five indicators of ‘0’ in the diagonal for the five SKIVE elements themselves as well as  $\binom{4}{2} = 6$  unique entries of ‘1’ for each pair of these four skill elements. Note also that adjacency matrices are symmetric, representing the occurrence of pairs of elements:

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}.$$

The adjacency matrices are then summed across time points to create representations of the epistemic frame as it develops over time. Summary measures such as the *relative centrality of*

*each node in the network* and the *overall density of the network* at any given point in time can be evaluated by computing properties of the matrices, either using existing algorithms from social network analysis, or one of several developed specifically for networks that have properties characteristic of ENA networks (see Shaffer et al, in press).

To facilitate communication about developing epistemic frames, the association structure of the variables that represent the epistemic frame elements can be viewed as an *undirected network with multiple nodes*. These multidimensional networks can be visually represented by projecting them onto a two-dimensional space as a *dynamic network graph*, generated according to the Kamada-Kawai algorithm (1989; Shaffer et al., 2009). Such representations are typical for graphically representing association structures (i.e., adjacency matrices or social networks); circles represent the nodes of the network (i.e., SKIVE elements) and the distances between the nodes represent the strength of the association between nodes, defined according to entries in the adjacency matrix (Shaffer et al., 2009; see also Wasserman & Faust (1994) for a review of dynamic network graphs and social network theory).

For example, Figure 4 shows the state of a network for a single learner at three different points during the game play (a video of the sequence of frame development is available at [www.youtube.com/watch?v=pctE4uXimFw](http://www.youtube.com/watch?v=pctE4uXimFw)). Different colors represent different frame elements and multiple components (e.g., multiple skills) are coded for each element. The graph shows that the network becomes denser over the course of the game play and that certain skills and types of knowledge become more central to the developing epistemic frame over time. Part of our future research effort is being directed toward developing appropriate visualizations to represent the salient properties of ENA analyses in graphic form.

Moreover, to describe and evaluate developing expertise using network graphs, it is important to have statistics that allow for comparisons of epistemic frame representations of (a) a single learner at different points in time, (b) different groups of learners at the same point in time, and (c) different groups of learners at different point in time. Moreover, networks of learners need to be compared to idealized network states or network states from professional experts in the field.

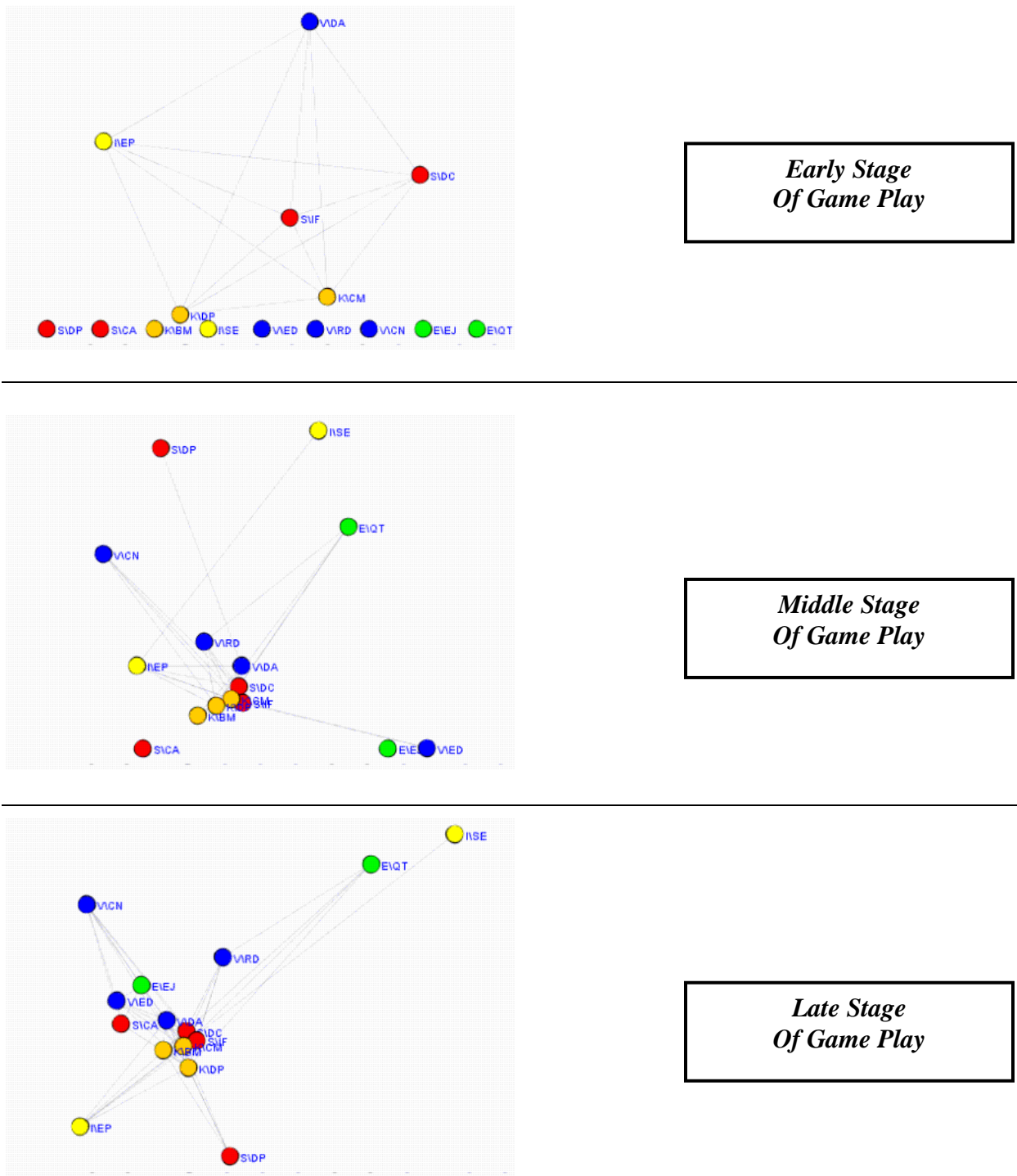


Figure 4 A developing epistemic frame at early, middle, and late stages of game play.

### *Assembly Model Decisions*

The decisions that have to be made within the three ECD models discussed so far show that the sequencing of different tasks within an epistemic game is critical to collect data that provide reliable information about individual learners or groups of learners. In a pragmatic sense, sequencing decisions for tasks are driven by the structure of the real-life practica that epistemic games mimic. Underlying the design of the real-life practica is a model of learning or developing expertise in the domain, which should be reflected in the different tasks (Bagley & Shaffer, 2009). Specifically, tasks need to be created and sequenced in such a manner that they provide learners the appropriate scaffolding for developing the needed expertise but also differentiate, to some degree, between learners with different degrees of expertise.

For example, if the objective is to measure the quality of oral communication skills for learners, it is important to design a task that provides opportunities for displaying communication skills of different levels of quality. Having learners simply read aloud the results of a preference survey would not be sufficient to elicit oratory performances of different qualities – in fact, any performance differences would reside in the written product. A summary presentation of a project proposal would seem like a more appropriate candidate task for differentiating presentation skills. Yet, summary presentations also differ in the range of performances they elicit depending on, for instance, the range of argumentative complexity and integration of different resources within the presentation.

The task and activity sequence in *Urban Science* is represented through a *frameboard*, which describes, in 15-minute segments, the flow of the game. It specifically lists:

1. the task that is currently performed,
2. which aspects of the task are performed by individual learners or groups of learners,
3. the nature of the interactions that learners have with mentors and the kinds of questions mentors are supposed to be asking to elicit particular kinds of evidence,
4. the nature of the interactions that learners have with virtual characters and the information that is provided to them, and
5. for each frame element, which particular component is required for task completion, what kind of evidence is expected, and where in the process or product data the evidence can be located.

Thus, like a *table of specifications* for more familiar assessments, a frameboard makes explicit linkages among the different ECD models. It can serve as a concise basis for building statistical models and summarizing their output into a coherent narrative about learners' developing epistemic frames. It also forms the basis for describing re-usable task templates that could be combined in different ways in new versions of the game or other epistemic games that focus on urban planning. Of particular importance for our future research will be the identification of situational features that can be encoded into task model variables and to link these variables explicitly to elements of the statistical models that we use.

#### *Presentation Model Decisions*

After deciding how different assessment tasks need to be sequenced in an epistemic game, the following decisions need to be made: (a) how requisite information for the tasks is provided, (b) how reflections of individual learners are recorded, (c) how interactions between learners are recorded, and (d) how products are recorded. From a technical perspective, recording interactions electronically, either in written or in audio form, has advantages in that it eliminates the need for time-consuming transcriptions. Deciding on the means of communication has substantive as well as technical implications, however. From a substantive perspective, learners need to be made familiar with each particular interface that is used to level out inter-individual differences in information technology literacy. This may also change the nature of the data that are collected. For example, communication via instant messaging will lead to a different flow of communication than real-life communication and the breadth and depth of topics covered is constrained by the typing speed of individuals. In the case of second or foreign language learners, written communicative ability in this genre will also affect the flow of interaction and reasoning as it can be strikingly different from their oral communicative ability. These are, of course, empirical questions that await multi-method inquiry. Qualitative and statistical methods can help determine whether such effects do, in fact, occur, and if so what impact they have on learners' experiences and on the resulting data collected.

The presentation of tasks interacts with the sequencing of tasks, which impacts at what point in time different kinds of information can be collected about learners. For example, if mentors are physically present on location they can accommodate learners in different environments; resulting conversations might reveal useful information about the learners' epistemic frames when acting in these environments. If mentors are in distant locations, learners would have to be

provided with wireless electronic access to them via laptops or handheld devices and may not have access to them at all times. Again, the extent to which this impacts either game play or frame analysis is an empirical issue and is part of our concerted research efforts for epistemic games.

In the current version of *Urban Science* all tasks and most communication are accessed through a unified game platform, which is essentially a web portal for the fictitious company *Urban Design Associates*. The learners receive e-mails from virtual characters that are represented by photos of real people. The e-mails describe the tasks that they have to perform and the resources that are made available to them via the platform. The tasks are a mixture of real-world tasks (e.g., learners travel to the actual neighbourhoods in Madison to take photos and notes) and virtual-world tasks (e.g., learners research information about background reports and write such a report).

As stated in the task model section, interactions between mentors and learners are currently conducted in real-world settings (i.e., at the site where the game is played) but will be completely electronic (i.e., via e-mail and instant messaging) in future versions of the game. Future research will investigate whether critical information about learners' use of their epistemic frame elements is lost when communication is conducted completely electronically in terms of (a) what learners write, (b) how much they write, and (c) how frequently they write.

#### Validation Research for Epistemic Games

All of the questions and research objectives that we have raised so far concern validation research in some form or another. In order to frame such work not solely from the perspective of different ECD framework models, we will now briefly revisit the different validity aspects identified by Messick (1995) that we introduced in the first main section of the paper. We do this to show how validation frameworks and the ECD framework can go together hand in hand to lay out a comprehensive research programme for epistemic games.

The relative importance of gathering evidence for each of these validity aspects in an epistemic game context is clearly different from a traditional proficiency assessment for admission, placement, or certification purposes. The major emphasis of an epistemic game is on developing expertise over the course of a well-coordinated learning experience and, thus, on diagnostic formative reporting of the epistemic frame development in a low-stakes environment.

Some of the research that is laid out below is currently being conducted with *Urban Science* while other research is envisioned for the future.

#### *Content Validity*

To address whether the kinds of tasks that are asked of learners in an epistemic game match tasks that are relevant to educational experiences in urban planning, detailed ethnographic studies of the planning practicum on which the game is based need to be conducted (see Bagley & Shaffer, 2009). Ideally, ethnographic analyses of practicum experiences of learners would also use ENA as an analytic method and, thus, provide a directly comparable baseline from which to assess learners' frame development in the relevant epistemic game; such an approach is currently taken for a different epistemic game *science.net* ([www.epistemicgames.org/eg](http://www.epistemicgames.org/eg)).

Results from such studies should also be cross-validated with ethnographic studies of practicing professionals unless the evidence for the fidelity between the practicum structure and professional practice is strong. Among other things, such studies need to analyze the flow of the curriculum, how tasks are tailored to different learner groups, how learner groups are supported to develop expertise, the roles that different learners play in different activities, and how learner groups are assessed. The principal outcome of ethnographic studies based on ENA would be a more finely-tuned theory of how the developing epistemic frames of novice and expert urban planners can be characterized and how this is reflected in the design, sequencing, implementation, and analysis of tasks throughout the practicum.

#### *Substantive Validity*

A focus on the mental processes that learners engage in when they respond to assessment tasks, which is the traditional focus of this aspect of validity, is strongly dominated by an information-processing perspective in applied cognitive psychology (e.g., Mislevy, 2006; see Rupp & Mislevy, 2007). In epistemic games, the dominant perspective is, perhaps, more aptly characterized as socio-cognitive or socio-cultural as much of the relevant learning happens through collaboration and cooperation. Hence, merely asking learners to *think-aloud* with a focus on their mental operations (e.g., Leighton, 2004; Leighton & Gierl, 2007), either during or after relevant game episodes, seems to bracket a large part of their relevant experiences as it might focus too much on the application of skills and knowledge. Accordingly, process data and product data collection have to include prompts for learners to reflect on their understanding of

values, identity, and epistemology as it develops and on how certain activities and interactions helped them to shape it.

### *Structural Validity*

The issue of scoring is a very challenging one to tackle in epistemic games as learners are not currently provided with scores that serve as indicators for developing expertise. Of course, there is substantial un-scored feedback that they receive; more, in fact, than in many non-game learning situations. While a heavier weighting of open-ended tasks in a traditional assessment might reflect the assumption that more complex skills assessed by such tasks are more critical to successful real-world performance, essentially all tasks in an epistemic game are complex. However, an analysis of the time allotted for each task and subtask as well as the sequencing of the tasks along with their design characteristics can provide insight into which epistemic frame elements are considered most crucial. Since epistemic games are developed based on professional practica, evidence for structural validity is gathered through an indirect logical chain as described earlier. That is, if (a) the fidelity of the epistemic game is high vis-à-vis the professional practicum, if (b) the fidelity of the professional practicum is high vis-à-vis professional practice, if (c) the scoring of performance in the practicum and game are consistent, and if (d) the weights that scores assign and their summaries are consistent with the importance of epistemic frame element use in professional practice, then there is some baseline evidence for structural fidelity.

### *Predictive Validity*

Assessing the degree to which the experience of playing an epistemic game is predictive of a real-world outcome is challenging because relevant concurrent and predictive criterion measures are needed. Currently, evidence for predictive validity is provided by pre-intervention, post-intervention, and follow-up interviews that include critiques of real-life urban redevelopment plans and other related outcome measures. The initial instruments were validated in small expert-novice studies and work is currently underway to gather evidence for the reliability of those scores with a larger sample. More readily available measures for learners might be course grades, but such a suitable composite would have to be used to adequately reflect the range of SKIVE elements that a game like *Urban Science* activates and develops.

### *External Validity*

External validity is empirically anchored in multi-trait multi-method designs (Campbell & Fiske, 1959). Collecting evidence for external validity of the epistemic frame representations from epistemic games, no matter how these are constructed, would require that learners be assessed on a broad range of indicators. Some of these indicators would have to be for latent characteristics that are related to the epistemic frame elements and some of them would have to be for latent characteristics that are relatively unrelated. What makes this process challenging is that there are typically no reliable alternative assessments that measure similar latent characteristics, because epistemic games are developed to fill just this gap.

For example, while there are some assessments that measure individual skill sets or knowledge sets, there are probably few reliable indicators of values, identity, and epistemology characteristics for urban planners relevant to *Urban Science*. A suitable set of indicators could perhaps be constructed with items from the civics components of NAEP. However, it is an open question at this point what the magnitude of correlations would be that one would expect between parameters from a statistical model that creates an epistemic frame representation within an epistemic game and the latent-variable scores from a NAEP assessment.

### *Generalizability*

The question of generalizability is, in some form or another, always relevant for any assessment, including epistemic games. The types of generalizability evidence required depends on the purpose to which the epistemic game is put. In general, collecting data for generalizability evidence requires first and foremost a thoughtful experimental design. For example, it may be reasonable to ask whether alternative versions of the games, with tasks that are constructed to be comparable to each other, evidence similar statistical functioning. This would require that a designed experiment be set up in which learners are first matched on relevant background variables and then are randomly assigned to the different versions of the game.

A comparison of the resulting epistemic networks of the two game versions for the matched learners could then be conducted and averaged across learners within each condition. Similarly, the data that are currently collected on the use of e-mail and instant messaging for *Urban Science* provides some evidence of whether the epistemic frame representations are robust across different modes of data collection. Related to these data, an investigation of different rating

designs for the process and product data could also be implemented to explore the degree to which the results from such rating schemes are comparable.

### *Consequential Validity*

Gathering evidence of the impact of intended and unintended consequences, both positive and negative, that arise for the learners, alternative stakeholders, and the disciplines that are touched by the epistemic games requires long-term follow-up investigation. As far as learners and related stakeholders such as parents and teachers are concerned, it would make sense to conduct semi-structured interviews at pre-specified time intervals. For example, as stated earlier, in *Urban Science* learners are currently interviewed before and after the game, and again in follow-up interviews after 3-4 months about their experiences, beliefs, and knowledge states. This could be extended, both temporally and in terms of the scope of the interviews, but there are practical limits on the number of topics and prompts that can be included in an interview. Similarly, the impact that the epistemic game design and implementation has on larger societal perceptions about the utility of game-supported learning could be investigated. This could be done with critical analyses of informational reports in popular media and academic discourse surrounding peer-reviewed publications as is currently done in other gaming contexts (Nelson, Erlandson, & Denham, 2009).

### Conclusions

We have shown in this paper that we are currently at the beginning of an exciting new era for learning and assessment via epistemic games. Therefore, it is not surprising that we are at early stages of articulating what counts as defensible, trustworthy, and convincing evidence about certain empirically supported arguments. We are just beginning to learn what the evidentiary frameworks and belief systems of different stakeholder groups are who are touched in some way by epistemic game learning. On the statistical front we are just beginning to tackle the specific and unique complexities that data from epistemic games produce and are carefully opening the door into the realm of understanding how analytic methods such as ENA or full-fledged latent-variable models can be adapted to the epistemic games context.

While some evidence already exists that epistemic games produce measurable and meaningful change in the epistemic frames of learners, more work needs to be done to develop appropriate narratives about the validity of claims arising from epistemic game play to warrant that the change we see is change that we can deeply believe in. These narratives will need to be

grounded in quantitative and qualitative research traditions, which open possibilities for truly enriching interdisciplinary research. Epistemic games are innovative assessment of, for, and as learning are currently pushing the methodological boundaries of educational assessment (see O'Reilly & Sheehan, 2008, for a large-scale analogue in the area of reading). It is up to the educational measurement community as well as the learning sciences community to leverage these possibilities as well as to structure and qualify the surrounding discourse and its reception. Through our own research and those of colleagues we seek to contribute to creating a strong research programme for epistemic games that can contribute valuable information for debates about how 21<sup>st</sup> century skills can be assessed using innovative digital technologies. In this spirit, we are looking very forward to a continual cross-disciplinary intellectual engagement of diverse groups of specialists and practitioners who are passionate about learning and assessment in these exciting contexts.

Author Note

The work at the University of Maryland was made possible, in part, by a grant from the Support Program for Advancing Research and Collaboration (SPARC) within the College of Education at the University of Maryland awarded to the first author as well as, in part, by a grant from the John D. And Catherine T. MacArthur foundation awarded to Arizona State University (07-90185-000-HCD) and two grants from the National Science Foundation awarded to the University of Wisconsin at Madison (DRL-0918409 and DRL-0946372). The work at the University of Wisconsin at Madison was made possible, in part, by a grant from the John D. and Catherine T. MacArthur foundation as well as four grants from the National Science Foundation (DRL-0347000, DUE-0919347, DRL-0918409, and DRL-0946372). The opinions, findings, and conclusions or recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies, cooperating institutions, or other individuals.

## References

- Adams, R. J. (2006, April). *Reliability and item response modeling: Myths, observations and applications*. Presented at the 13<sup>th</sup> International Objective Measurement Workshop, Berkeley, CA, April 5-7.
- Allen, M. J., & Yen, W. M. (2002). *Introduction to Measurement Theory*. Long Grove, IL: Waveland Press.
- Almond, R. G., Williamson, D. M., Mislevy, R. J., & Yan, D. (in press). Bayes nets in educational assessment. New York: Springer.
- Almond, R. G., Mulder, J., Hemat, L. A., & Yan, D. (2006). *Bayesian network models for local dependence among observable outcome variables* (RR-06-36). Princeton, NJ: Educational Testing Service.
- American Psychological Association, National Council on Measurement in Education, & American Educational Research Association (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Bauer, D. J. (2007). Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research*, 42, 757-786.
- 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, E. L., Dickieson, J., Wulfeck, W., & O'Neil, H. F. (Eds.) (2007). *Assessment of problem solving using simulations*. Mahwah, NJ: Erlbaum.
- Bennett, W. Jr., Lance, C. E., & Woehr, D. J. (Eds.) (2006). *Performance measurement: Current perspectives and future challenges*. Mahwah, NJ: Erlbaum.
- Borsboom, D., Mellenbergh, G. J. (2007). Test validity in cognitive assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 85-118). Cambridge: Cambridge University Press.
- Brecht, J., Cheng, B., Mislevy, R., Haertel, G., & Haynie, K. (2009). *The PADI System as a complex of epistemic forms and games* (PADI Technical Report 21). Menlo Park, CA: SRI International.

- Brennan, R. L. (2001a). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, 38(4), 295-317.
- Brennan, R. L. (2001b). *Generalizability theory*. New York: Springer.
- Campbell, D. T., & Fiske, D. W. (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- Clark, D., Nelson, B., Sengupta, P., & D'Angelo, C. (2009, September). *Rethinking science learning through digital games and simulations: Genres, examples, and evidence*. Presented at the Learning science: Computer games, simulations, and education workshop sponsored by the National Academy of Sciences, Washington, D.C.
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.
- de Boeck, P. & Wilson, M. (eds.) (2004). *Explanatory item response models: A generalized linear and non linear approach*. New York: Springer.
- Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Mahwah, NJ: Erlbaum.
- Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modelling: Concepts, issues, and applications*. Mahwah, NJ: Erlbaum.
- Ehrlich, T. (2000). *Civic responsibility and higher education*. Phoenix, AZ: American Council on Education and The Oryx Press.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (2008). *Longitudinal data analysis: A handbook of modern statistical methods*. New York: Chapman & Hall/CRC.
- Frisbie, D. (1988). Reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practice*, 10, 55-65.
- Gee, J. P. (2003). *What video games have to teach us about learning and literacy*. New York: Palgrave / Macmillan.
- Gibson, D., Aldrich, C., & Prensky, M. (Eds.). (2006). *Games and simulations in online learning: Research and development frameworks*. Hershey, PA: Information Science Publishing.

- Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, 33, 204-229.
- Hale, J. (1999). *Performance-based certification: How to design a valid, defensible, and cost-effective program*. New York: Pfeiffer.
- Joint Committee on Testing Practices (2004). *Code of fair testing practices in education*. Retrieved on November 11, 2008 from <http://www.apa.org/science/fairtestcode.html>
- Kamada, T. & Kawai S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31, 7-15.
- Kane, M. (2007). Validation. In R. J. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 18-64). Westport, CT: Praeger.
- Kane, M. (2008, October). *Content-based interpretations of test scores*. Presented at the 9<sup>th</sup> annual Maryland Assessment Conference entitled The Concept of Validity: Revisions, New Directions and Applications at the University of Maryland, College Park, MD, October 9-10.
- Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.
- Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice*, 23(4), 6-15.
- Leighton, J. P., & Gierl, M. (2007). Verbal reports as data for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 146-172). Cambridge: Cambridge University Press.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (4<sup>th</sup> ed.) (pp. 13-103). New York: Macmillan.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749.
- Mislevy, R.J. (2004). Can there be reliability without “reliability”? *Journal of Educational and Behavioral Statistics*, 29, 241-244.
- Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4<sup>th</sup> edition) (pp. 257-305). Portsmouth, NH: Greenwood Publishing Group.

- Mislevy, R. J. (2008, October). *Validity and narrative robustness*. Presented at the 9<sup>th</sup> annual Maryland Assessment Conference entitled The Concept of Validity: Revisions, New Directions and Applications at the University of Maryland, College Park, MD, October 9-10.
- Mislevy, R. J., & Braun, H. I. (2003, May). *Intuitive test theory*. Presented at the Annual Dinner Meeting of the Princeton Association for Computing Machinery (ACM) and Institute of Electrical and Electronics Engineers (IEEE) Computer Society Chapters, Kingston, NJ, May 22.
- Mislevy, R. J., Almond, R. G., & Steinberg, L. S. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-62.
- Mislevy, R. J., Steinberg, L. S., Almond, R. G., & Lukas, J. F. (2006). Concepts, terminology, and basic models of evidence-centered design. In D. M. Williamson, I. I. Bejar, & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 15-48). Mahwah, NJ: Erlbaum.
- Moss, P. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5-12.
- National Assessment Governing Board (2006). *Civics framework for the 2006 National Assessment of Educational Progress*. Retrieved November 11, 2008 from [http://www.nagb.org/civics\\_06.pdf](http://www.nagb.org/civics_06.pdf)
- Nelson, B. C., Erlandson, B., & Denham, A. (2009, April). *A Design View of Assessment in Complex Game Environments*. Presented at the third meeting of the Assessment of 21<sup>st</sup> Century Skills Working Group sponsored by the MacArthur Foundation, Tempe, AZ, April 14-16.
- O'Reilly, T., & Sheehan, K. M. (2008). *Cognitively based assessment of, for, and as learning: A 21<sup>st</sup> century approach for assessing reading competency* (RR-09-04). Princeton, NJ: Educational Testing Service.
- Partnership for 21<sup>st</sup> Century Skills (2008). *21<sup>st</sup> century skills, education, and competitiveness: A resource and policy guide*. Tuscon, AZ. Available online at [www.21stcenturykills.org](http://www.21stcenturykills.org)
- Phelan, M. J. (1990). Estimating the transition probabilities from censored Markov renewal processes. *Statistics & Probability Letters, 10*, 43-47.
- Plass, J. L., Homer, B. D., & Hayward, E. (2009). Design factors for educationally effective animations and simulations. *Journal of Computing in Higher Education, 21*(1), 31-61.

- Rupp, A. A. (2008, April). *Psychological vs. psychometric dimensionality in diagnostic language assessments: Challenges for integrated assessment systems*. Invited presentation at the conference entitled Reading in the 21<sup>st</sup> Century co-sponsored by ETS and IES, Philadelphia, PA, April 16-19.
- Rupp, A. A., & Mislevy, R. J. (2007). Cognitive foundations of structured item response theory models. In J. Leighton & M. J. Gierl (Eds.), *Cognitively diagnostic assessment for education: Theory and applications* (pp. 205 - 241). Cambridge: Cambridge University Press.
- Rupp, A. A., & Templin, J. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, 6, 219-262.
- Rupp, A. A., Templin, J., & Henson, R. (in press). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford Press.
- Rupp, A. A., Choi, Y.-Y., Gushta, M., Mislevy, R. J., Thies, M.-C., Bagley, E., Hatfield, D., Nash, P., & Shaffer, D. (2009a, May). *Statistical research for data structures from epistemic games: A brainstorm of ideas*. Presented at the third meeting of the Assessment Group for Sociocultural and Games-based Learning sponsored by the John D. and Catherine T. MacArthur foundation, Phoenix, AZ, May 14-17.
- Rupp, A. A., Choi, Y., Gushta, M., Mislevy, R. J., Bagley, E., Nash, P., Hatfield, D., Svarowski, G., & Shaffer, D. (2009b). Modeling learning progressions in epistemic games with epistemic network analysis: Principles for data analysis and generation. *Proceedings from the Learning Progressions in Science Conference* held in Iowa City, IA, June 24-26.
- Seeratan, K., & Mislevy, R. (2009). *Design patterns for assessing internal knowledge representations* (PADI Technical Report 22). Menlo Park, CA: SRI International.
- Shaffer, D. W. (2006a). *How computer games help children learn*. New York: Palgrave / Macmillan.
- Shaffer, D. W. (2006b). Epistemic frames for epistemic games. *Computers and Education*, 46(3), 223-234.
- Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., Franke, K., Rupp, A. A., & Mislevy, R. J. (2009). Epistemic network analysis: A prototype for 21<sup>st</sup> century assessment of learning. *International Journal of Learning Media*, 1(2), 33-53.
- Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78, 153-189.

- Shute, V. J., & Spector, J. M. (2008). *SCORM 2.0 white paper: Stealth assessment in virtual worlds*. Unpublished manuscript. Retrieved November 14, 2009 from <http://www.adlnet.gov/Technologies/Evaluation/Library/Additional%20Resources/LETSI%20White%20Papers/Shute%20-%20Stealth%20Assessment%20in%20Virtual%20Worlds.pdf>
- Shute, V. J., Dennen, V. P., Kim, Y.-J., Donmez, O., & Wang, C.-Y. (in press). 21<sup>st</sup> century assessment to promote 21<sup>st</sup> century learning: The benefits of blinking. In J. Gee (Ed.), *Games, learning, assessment*. Boston, MA: MIT Press.
- Shute, V. J., Ventura, M. Bauer, M., and Zapata-Rivera, D. (2008). *Monitoring and fostering learning through games and embedded assessments* (Research report RR-08-69). Princeton, NJ: Educational Testing Service.
- Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, 26(4), 21-28.
- Storey, J. K., Kopp, K. J., Wiemer, K., Chipman, P., & Graesser, A. C. (in press). Using AutoTutor to teach scientific critical thinking skills. *Behavior Research Methods*.
- Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20, 345-354.
- Templin, J., & Henson, R. (2009, April). *Practical issues in using diagnostic estimates: Measuring the reliability and validity of diagnostic estimates*. Presented at the annual meeting of the National Council on Measurement in Education (NCME), San Diego, CA.
- Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.
- Wasserman, S. & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.
- Wild, C. L., & Rawasmany, R. (Eds.). (2007). *Improving testing: Applying process tools and techniques to assure quality*. Mahwah, NJ: Erlbaum.
- Williamson, D. M., Mislevy, R. J., & Bejar, I. (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Erlbaum.

## Appendix: Urban Science, an Example of an Epistemic Game

### *Objective of Game*

In the *Urban Science* game learners assume the role of urban planners in redesigning neighborhoods in the city of Madison, WI where the development team is geographically located. Versions that expand this geographic scope are currently under development. In *Urban Science*, learners must use information, tools, and methods typically used by urban planning professionals. For example, learners collect neighbourhood information that is provided to them by virtual characters from stakeholder groups, they inspect the real-life neighbourhoods through site visits, they integrate different pieces of information via a virtual interface that overlays zoning information onto a geographical map of the neighbourhood, and summarize their plans in a cumulating report and presentation. During the course of the game, learners work individually and interact with others, which include other learners as well as mentors that guide them through the game. Interaction is currently conducted in real-life settings (i.e., meetings) but will be take place via online means (e.g., e-mail or instant messaging) in the future.

### *Game Interface*

To provide a sense of the game interface, Figure A1 shows a screenshot of the main screen of the game with an email that the learner received from a virtual character. In the email, a specific task is described - here creating a bio and posting it – and resources for completing the task are made available to the learner upon reading the email. The screen also shows links to the learner's inbox, the planning notebook for tracing works in progress, and the different projects that the learner has worked on. Depending on the version of the game and the time available for its implementation, the game consists of a sequence of four broad tasks that ask learners to develop plans for re-zoning either one or multiple different neighbourhoods in Madison, WI. As shown through the links on the right side of Figure A1, the neighbourhoods in the particular version of the game shown here are *State Street*, *Schenk-Atwood*, *Northside*, and *Madison East*.

### *Task Structure*

All tasks are similar to one another in that learners are first split up into groups representing different stakeholders with the task of developing an argument for urban planning that highlights their particular stakeholder perspective. They are then re-grouped to develop a joint proposal for the redevelopment of each neighbourhood that incorporates all stakeholder group perspectives. The key steps in the redevelopment process are to translate relevant information into a preference

survey, which requires players to input and justify choices about zoning in an interactive simulation interface. Each game culminates in a few summary tasks that consist of (a) an issue statement, (b) a summary plan, and (c) a final presentation. For example, Figure A2 shows part of a final summary proposal for *Madison-East* by one learner group with learners representing different stakeholder groups. It shows a map of the neighbourhood along with the zoning indicators from the preference survey. Underneath the map is the first part of a longer segment of text that discusses the rationale for the choices that were made in redesigning this neighbourhood.



Figure A1 Screenshot of e-mail in the main screen of the *Urban Science* interface.

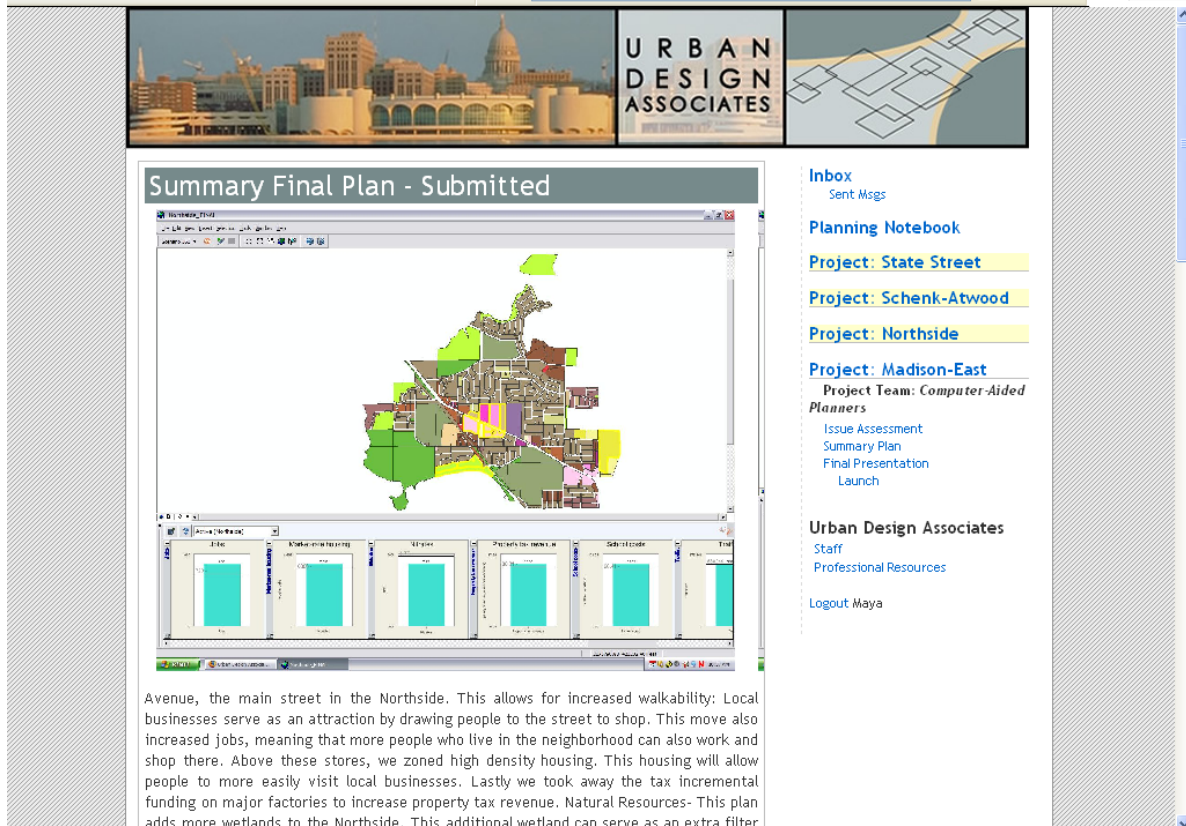


Figure A2 Screenshot of first part of redevelopment plan for *Madison-East*.