

Running head: INVESTIGATION OF NOVEL METHOD FOR EPISTEMIC GAMES

**Modeling Learning Trajectories with Epistemic Network Analysis: An Investigation of a Novel Analytic Method for Learning Progressions in Epistemic Games**

Younyoung Choi, André A. Rupp, Matthew Gushta, & Shauna J. Sweet  
EDMS Department, University of Maryland

Contact information for paper:

André A. Rupp  
EDMS Department, University of Maryland  
1230-A Benjamin Building  
College Park, MD 20742  
Phone: (301) 328-0747  
E-mail: [ruppandr@umd.edu](mailto:ruppandr@umd.edu)

**Abstract**

*Epistemic games* are designed to help players develop domain-specific expertise that characterizes how professionals in a particular domain reason, communicate, and act (Bagley & Shaffer, 2009; Shaffer, 2006b). To analyze the complex data that arise from these games, a novel analytic method grounded in social network analysis called *epistemic network analysis* (ENA) has been recently proposed (Rupp, Gushta, Mislevy, & Shaffer, 2010; Rupp et al., 2009; Shaffer et al., 2009). In this paper, we introduce the basic ideas of this method and report on the preliminary results of an ongoing research program that investigates whether ENA statistics are sensitive to detecting players' differential learning trajectories throughout different game structures under different solution strategies. Preliminary results show a complex emerging picture of the conditions under which one ENA statistic can be suitable for this purpose.

## Introduction

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 (Bagley & Shaffer, 2009; Partnership for 21<sup>st</sup> Century Skills, 2008; Shute, Dennen, 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, interpersonal communication and collaboration skills, digital networking and operation skills, intra- and intercultural awareness and identity, and cross-cultural sensibility. 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 the 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).

*Epistemic games* have been developed in recent years to help players develop domain-specific expertise that characterizes how professionals in a particular domain reason, communicate, and act (Bagley & Shaffer, 2009; Shaffer, 2006a). Although there are many games- and simulation-based opportunities for transforming practices, perceptions, and commitments regarding learning in the 21<sup>st</sup> century (e.g., Gee, 2003; Gibbons, Aldrich, & Prensky, 2006), epistemic games are explicitly based on theory of learning in the digital age and are designed to allow learners to develop domain-specific expertise under realistic constraints. For example, learners may learn what it is like to think and act like journalists, artists, business managers, or engineers by using digital learning technologies to solve realistic complex

performance tasks. 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. To illustrate these ideas, Appendix B shows illustrative screenshots with short explanations for one particular epistemic game called *Urban Science*, which was modelled after core practices in the field of urban planning.

As one might expect, modern latent variable measurement models originally designed for traditional large-scale assessments with relatively constrained tasks, struggle to accommodate the contextual dependencies and more open-ended learner performances characteristic of epistemic games. Moreover, implementations of epistemic games are typically small in scale and the statistical demands of most latent variable models for reliable parameter estimation are thus too strong for them to represent viable analytic approaches. *Bayesian inference networks*, for example, typically require several hundred or thousands of learners and / or very strong theoretical beliefs in the form of empirical priors for a reliable calibration (e.g., Shute, Dennen, Donmez, & Wang, in press; West et al., 2009). Thus, there are currently no off-the-shelf statistical models that can be applied directly to epistemic games to satisfy the desired scaling and reporting purposes; alternative modeling approaches grounded in non-parametric methods appear to be more promising in this regard.

In this paper, we provide a brief introduction to a novel analytic method for epistemic games and report on preliminary results from an ongoing research program designed to investigate the method. The method in question is called *epistemic network analysis* (ENA) (Rupp, Gushta, Mislevy, & Shaffer, 2010; Rupp et al., 2009; Shaffer et al., 2009) and is a relatively straightforward adaptation of simple computational algorithms and associated descriptive statistics from social network analysis. Contrary to social network analysis, however, the focus of ENA is not on constellations of individual people whose interaction patterns need to be mapped, but on the latent characteristics of individuals whose association structure needs to be mapped.

Thus, similar to fully parameterized latent variable models, ENA creates statistical representations of the latent characteristics of learners that are of interest with a specific longitudinal focus on how the association between these characteristics changes over time. ENA

has been applied to real data collected in several different epistemic games but as of yet has not been thoroughly investigated using simulation studies that manipulate conditions representing a wide variety of realistic game-play scenarios. The simulation studies described in this paper are designed to fill this gap in games-based assessment at the interdisciplinary intersection of research in the learning sciences, human-computer interaction, and educational and psychological measurement.

Since the results in this paper are preliminary the objective is not to provide a comprehensive answer to how this method can perform. Instead, the purpose is to raise awareness of the need for creating and investigating alternative measurement approaches for serious educational games that can address the specific reporting needs for these contexts in the absence of off-the-shelf methods that can be directly applied to these contexts.

This paper is divided into four sections. First, we review the basic principles and procedures of ENA and how learners' progress through epistemic games are captured and coded; these descriptions are based on collaborative work with researchers at the University at Wisconsin at Madison who have developed these games and initially proposed this method. Second we describe the learner and task parameters used in the simulation study and the process by which that data was generated. In the third section we discuss the performance of a particular ENA statistic as a measure for differentiating between learners with different learning trajectories across different game contexts. We close this paper with an overview of related ongoing and future work.

### **Basic Principles and Procedures of Epistemic Network Analysis**

In epistemic games the sequence of activities in the game is divided into units, which are also called *slices*, *episodes*, or *activity segments*; we will use the term *slices* in this paper for simplicity. The slices could be defined in a sequential and contiguous fashion based on objective criteria such as time (e.g., 15-minute intervals) or macro-task boundaries (e.g., the beginning and end of a task), but they could also be defined in a sequentially but non-contiguous fashion based on objective criteria such as interactional structures (e.g., pair work versus large-group discussions) or more subjective criteria such as thematic foci (e.g., content analysis of interaction segments).

Once the sequence of activities in the game has been segmented the objective is to extract relevant information about the epistemic frames of the learners from observable products. These

products can contain sequences of actions (e.g., sequences of mouse clicks, residing times in segments of the interface), learner products (e.g., report drafts, presentation drafts), or discourse segments (e.g., discussions between pairs of learners, questions posed to mentors). Once the relevant observable products have been identified, the products are coded with respect to relevant evidence about the epistemic frame elements activated by the game and utilized in the process of task completion. By no means are learners required to use every epistemic element all of the time; rather, tasks require the successful application of only particular abilities and learners utilize different combinations of skills and abilities over the course of game play. The resulting codes constitute the response data for the game that is used as the input for ENA analyses, which transforms those codes into numerical and visual representations of learners' emerging expertise that can be fed back to the learners and their mentors during the game.

### ***Structure of Resulting Data***

While the number of epistemic frame elements on which evidence is collected at each slice depends on the grain size of the desired feedback, in this study we use five macro-categories to describe the content and structure of learners' emerging expertise during game play. We label these categories as *Skills* (S), *Knowledge* (K), *Identity* (I), *Values* (V), and *Epistemology* (E):

1. *Skills (S) (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 (K) (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 (I) (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 (V) (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 (E) (general, planning-specific)*: being able to understand what counts as relevant evidence that justifies actions as legitimate within the urban planning community.

Observable evidence for a learner's reliance on any one of these SKIVE elements in his or her epistemic frame is thus recorded by a '1' for a given slice. Though these data can technically be of any nature, our analyses are based on dichotomous (i.e., yes-no / 1-0) codes to indicate whether evidence for a particular epistemic frame element was present or absent. A sequence of

observations for a single learner over the course of a game might look like the one shown in Table 1.

Table 1  
*Sequence of Observations for Single Learner*

Slice	S	K	I	V	E
1	1	1	0	1	0
2	0	1	0	1	0
3	0	0	1	0	1
4	1	1	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮
<i>T</i>	1	1	0	1	1

From the raw data captured at each slice, an *adjacency matrix* is created, which is a statistical representation of the relational structure between the epistemic frame elements. Adjacency matrices contain entries of ‘1’ whenever two frame elements are used by a learner concurrently within a slice and ‘0’ otherwise; Table 2 shows a sample adjacency matrix for slice 1 for the learner whose data are shown in Table 1.

Table 2  
*Sample Adjacency Matrix for a Single Time Slice*

	S	K	I	V	E
S	0	1	0	1	0
K	1	0	0	1	0
I	0	0	0	0	0
V	1	1	0	0	0
E	0	0	0	0	0

*Note.* This adjacency matrix is for slice 1 in Table 1.

Since adjacency matrices are available for each slice, the evidence in them can be accumulated across different slices by simply summing the individual entries in the adjacency matrices across the slices of interest. This resulting matrix is called a *cumulative adjacency matrix*; Table 3 shows such a matrix for the first four slices of the learner from Table 1.

Table 3  
Cumulative Adjacency Matrix

	S	K	I	V	E
S	0	2	1	2	1
K	2	0	1	3	1
I	1	1	0	1	1
V	2	3	1	0	1
E	1	1	1	1	0

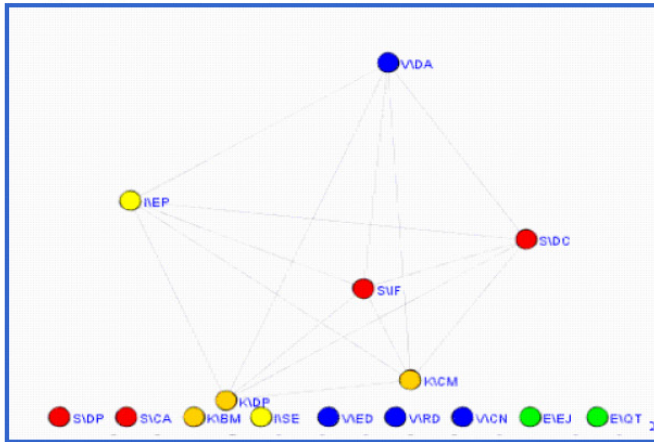
Consistent with the theoretical emphasis on interconnectedness between epistemic frame elements, this particular process of coding and accumulation captures only the co-occurrence of SKIVE elements. Evidence for individual SKIVE elements, in the absence of other SKIVE elements, is currently discarded, even though it could easily be captured and incorporated into the analyses.

In practice, cumulative adjacency matrices across time slices are then projected into a two-dimensional representation that shows the interconnections between the SKIVE elements as nodes in a network. Again, different projection algorithms exist; the Kamada-Kawaii algorithm has proved useful in these initial stages of work. Figure 1 on the next page, which is taken from Rupp, Gushta, Mislevy, and Shaffer (2010), shows such network representations at three different stages in the game for a particular learner. In other words, ENA is the numerical method for transforming observed responses to time slices and the two-dimensional network graphs are the visual representations of the key features of the data patterns from an ENA perspective; both of these representations are mappings of the theoretical construct relationships onto the empirical statistical realm.

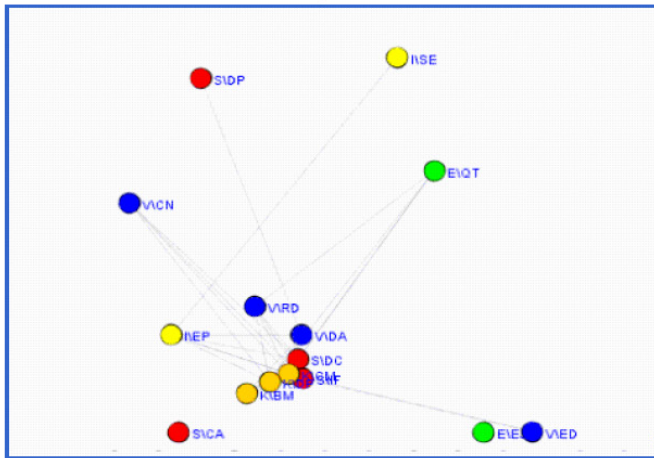
### ***Weighted Density***

There are a variety of statistics that are available for ENA and our research program is continually exploring how different principles from multivariate data analysis can be applied to the work with ENA to improve its performance. To help organize discussions about ENA performance, one can distinguish existing ENA statistics in terms of whether they provide *global marginal information* (i.e., information about the network), univariate marginal information, or bivariate information. In this paper we focus only on one specific global statistic for ease of communication.

*Early Stage of Game Play*



*Middle Stage of Game Play*



*Late Stage of Game Play*

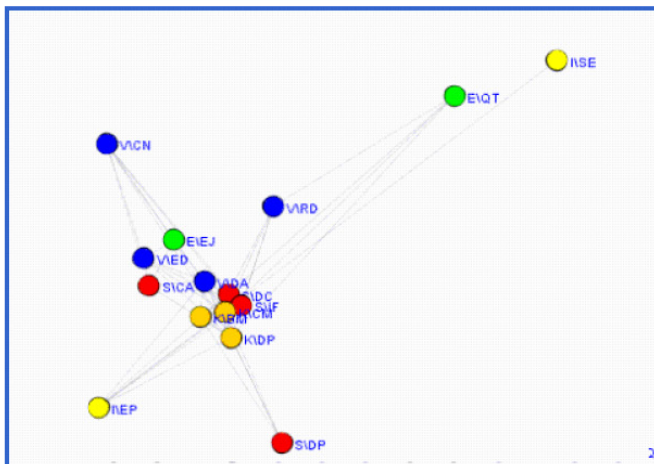


Figure 1 Epistemic network representations at different stages of game play.

In the context of ENA, *global marginal information* is any information that summarizes for one or multiple learners over multiple activities or slices of the game (i.e. rows of the data matrix) and also across SKIVE elements (i.e., columns of the data matrix). In terms of notation, we will use  $t = 1, \dots, T$  to refer to slices in reference to the fact that they prototypically denote time or task and because we need the letter  $s$  for a task parameter later on,  $f = 1, \dots, F$  to refer to epistemic frame elements, and  $N$  to denote the network (i.e., epistemic frame representation) of a particular learner.

The key global marginal statistic in ENA that we examine in this paper is the *overall weighted density* (WD) of the network for any given cumulative adjacency matrix, which is computed as follows:

$$WD_t(N) = \sqrt{\frac{1}{2} \sum_{f=1}^F \sum_{f'=1}^{F'} a_{ff',t}^2}$$

where  $a_{ff',t}^2$  is the squared entry in the cumulative adjacency matrix for nodes  $f$  and  $f'$  at slice  $t$ . Note that the total sum is divided by two because the cumulative adjacency matrix is symmetric. The overall weighted network density thus represents the average pair-wise association between nodes in the network that represents the epistemic frame.

### **Research Objectives and Their Motivation**

ENA is a simple computational method whose creation was borne out of practical necessities within learning contexts that use epistemic games. The research team wanted to provide feedback to learners and their mentors throughout the game play in a manner that supports reliable meaningful interpretations of learners' development. This was to be done in the context of data structures that have numerous layers of complexities that make the application of modern latent-variable models notoriously challenging.

Most importantly, the data are multivariate in nature and contain numerous contextual dependencies, both because learners solve complex performance tasks and because learners interact heavily with other learners in the process. The data are also longitudinal in structure, within the bounds of a single game, and interpretations are made predominantly about individual learners. Game segments that consist of multiple slices are also not necessarily parallel in nature so that envisioning blocks of game segments as “parallel” segments – akin to parallel tests in traditional assessment contexts – was not simple to do. Most important for data-analytic purposes

perhaps, the sample size available for model estimation is typically small with only about 15-20 learners playing these games in an after-school program or comparable educational context.

Therefore it was not possible to simply apply a complex latent-variable model to these data despite the fact that modern latent-variable models offer accommodations for these complexities. For the purposes of the studies described in this paper, the first step was to create systematic simulation studies to investigate empirically how well different statistics from ENA are able to differentiate between learners who play different kinds of games.

Since ENA is not a parametric model or even a fully formulated non- or semi-parametric model in the traditional sense – it is a computational method from multivariate data analysis and social network analysis - the purpose of the simulation study was not to assess parameter recovery of this method. Rather, the objective was to generate data according to a plausible mechanism grounded in sound measurement principles that could separate the influence of game task and learner parameters on observed responses.

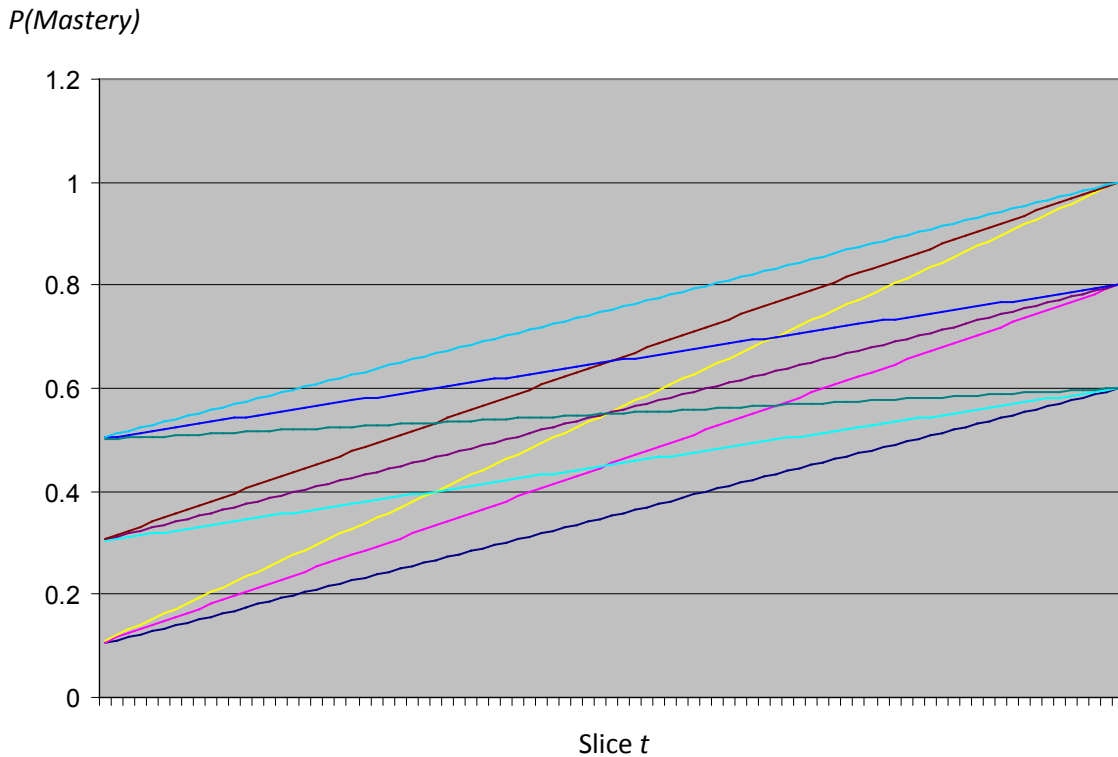
The measurement principles that we relied on were grounded in *item response theory* (IRT) (de Ayala, 2009; Embretson & Reise, 2000) and *diagnostic classification models* (DCMs) (Rupp & Templin, 2008; Rupp, Templin, & Henson, 2010). In other words, we used the principles of parameter separation in these models and the representation of learner characteristics via mastery profiles in DCMs to generate our data. We deemed this superior to alternative generation mechanisms that merely placed marginal (i.e., row-wise or column-wise) constraints occurrences of SKIVE elements in the observed data matrix without any separation of task and learner effects.

### ***Specification of Learner Parameters***

In order to represent different learning trajectories for the different SKIVE nodes in the network we specified different trajectories of the mastery probabilities for the individual SKIVE elements across the different slices that comprise a full game. We distinguished three core sets of trajectories, two of them consisting of linear growth trends and one of them consisting of curvilinear growth trends.

In the first set of linear growth trend, non-zero slopes were modeled with different intercepts corresponding to different initial probabilities of mastery at the beginning of game play. The slopes themselves were determined via linear interpolation by setting the desired final probabilities of mastery at the end of the game play. Crossing three different initial probabilities

of (.1, .3, .5) with three different final probabilities of (.6, .8, 1) led to nine different trajectories in this set. In the second set of linear growth trend, zero slopes were modeled that corresponded to trends representing no growth over the course of game play. We used 11 different probabilities corresponding to the set (0, .1, .2, ..., 1.0); Figure 2 shows all linear growth trends.



*Figure 2* Linear learning trajectories for mastery probabilities of learners.

In the set of curvilinear growth trends, slow initial learning followed by quick growth spurts in later stages of the game play and quick learning in early stages of game play followed by a flattened growth trend in later stages of game play were modeled; a total of nine growth trends were modeled. The first subset of growth trends was modeled by using exponential functions with powers 2, 4, and 8, while the second subset of growth trends was modeled by using exponential functions with powers .5, .25, and .125 for a total of six trends; Figure 3 shows the curvilinear growth trends.

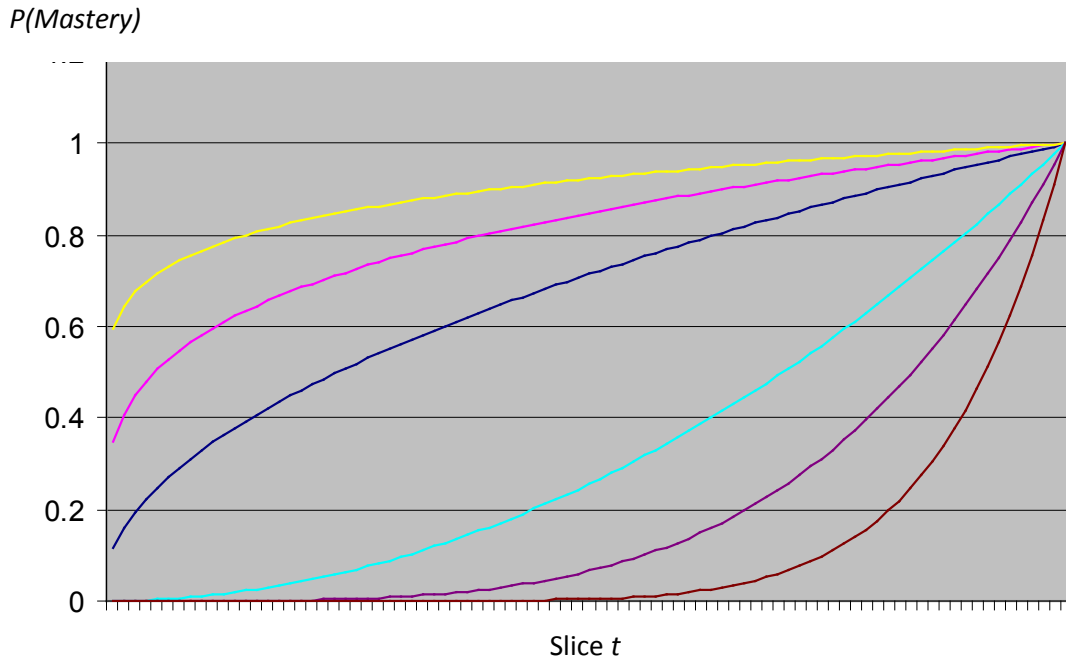


Figure 3 Curvilinear learning trajectories for mastery probabilities of learners.

Overall, we thus modeled a total of 26 growth trends, which provides a broad coverage of different trajectory patterns through the space of mastery probabilities for individual SKIVE nodes.

### ***Specification of Task Parameters***

In unidimensional and multidimensional IRT models task parameters operationalize operating characteristics such as difficulty, discrimination, and guessing or, more technically, location, slope, and asymptotes of item characteristic functions or surfaces. In DCMs the parameters that operationalize such concepts exist as well even though their meaning is slightly different due to the discrete nature of the latent variables and latent class structure. Since we are using the principles of DCMs specifically for this study, we will describe the task parameters in these models in a bit more detail.

In the parlance of traditional achievement tests, the two prototypical task parameters for DCMs are “slipping” and “guessing” parameters. Slipping parameters, typically denoted by the letter  $s$ , represent probabilities of responding inappropriately when learners have mastered a particular required skill or a set of required skills; thus, the reverse probability of  $(1 - s)$  is the probability of providing an appropriate response. Guessing parameters, typically denoted by the

letter  $g$ , represent probabilities of responding appropriately when learners have not mastered a particular required skill or a set of required skills. In other words,  $(1 - s)$  and  $g$  are the probabilities of responding appropriately when expected and when not as characterized by the mastery status of the learners.

In the context of epistemic game play, expertise is demonstrated through both the appropriate *application* and appropriate *suppression* of SKIVE elements. Given a particular reference pattern from an expert or mentor, learners that have high levels of expertise are expected to match the expert's or mentor's pattern as closely as possible, producing *efficacious solutions*. That is, learners are expected to produce solutions that are both effective (i.e., they solve the problem at hand) and efficient (i.e., they draw only on those frame elements necessary to solve the problem). Learners are expected to demonstrate their ability but to do so appropriately and - like experts - draw only on those epistemic frame elements relevant to a particular task. Under this conceptualization, observing a '0' in data string for a particular task is not necessarily evidence of a lack of ability; it may be evidence of emerging expertise in using that ability only as appropriate.

Contrary to traditional assessment contexts, it is not sufficient to characterize well-functioning tasks based only the likelihood that a learner will exhibit use of a particular SKIVE element as required. It is necessary to also set task parameters corresponding to the likelihood that a learner will suppress the use of such an element when it is outside the critical path of completing a given task. Given that a learner has mastered a particular epistemic frame element, she can "slip" by failing to demonstrate her ability as required by a particular task, or she can "slip" by failing to suppress that same ability when its use is inappropriate.

In both cases, a well-functioning task is one for which the probability of her slipping is low, but an appropriate parameterization needs to reflect that those features of a task that would make it less likely for a learner to demonstrate mastery as required by the task (difficulty) are qualitatively different than those task features (specificity) that would give rise to opportunities for a learner to utilize frame elements that are not critical to task completion. Similarly, a learner who has not yet mastered a particular element has two ways in which she can "guess" to match an expert's response pattern, and while the likelihood of her being able to produce a '1' should be low, not showing evidence of use of that epistemic frame element is far easier than showing evidence of it even if that is what the expert does.

To acknowledge this characteristic of epistemic games we distinguish between guessing and slipping parameters for each SKIVE node contingent on each of the possible expected expert responses, ‘0’ and ‘1.’ Defining task parameters in this way provides us greater flexibility and conceptual clarity by enabling us to characterize a task to reflect both task difficulty and specificity of task focus.

$$P(X = 1|\alpha = 1, E = 1) = 1 - s^{(1)} \quad (6)$$

$$P(X = 0|\alpha = 1, E = 0) = 1 - s^{(0)} \quad (7)$$

$$P(X = 1|\alpha = 0, E = 1) = g^{(1)} \quad (8)$$

$$P(X = 0|\alpha = 0, E = 0) = g^{(0)} \quad (9)$$

Using these parameter definitions we can now define the marginal probabilities of producing the desired response of ‘1’,  $P(X = 1|E = 1)$ , and of ‘0’,  $P(X = 0|E = 0)$ , for a particular SKIVE node. It is a function of both mastery states on the node,  $\alpha = 1$  (mastery) and  $\alpha = 0$ , or the respective probabilities of mastery,  $P(\alpha = 1)$ , and non-mastery,  $P(\alpha = 0)$ , as well as the task parameters as defined above:

$$\begin{aligned} P(X = 1|E = 1) &= P(X = 1|\alpha = 1, E = 1)P(\alpha = 1) + P(X = 1|\alpha = 0, E = 1)P(\alpha = 0) \\ &= 1 - s^{(1)}P(\alpha = 1) + g^{(1)}P(\alpha = 0) \end{aligned} \quad (10)$$

$$\begin{aligned} P(X = 0|E = 0) &= P(X = 0|\alpha = 1, E = 0)P(\alpha = 1) + P(X = 0|\alpha = 0, E = 0)P(\alpha = 0) \\ &= 1 - s^{(0)}P(\alpha = 1) + g^{(0)}P(\alpha = 0) \end{aligned} \quad (11)$$

The probabilities of mastery for each SKIVE node are given by the linear or curvilinear growth trends specified above while we set the task parameters deliberately with fixed values to describe an array of different game conditions. Table A1 shows the task parameter values for different conditions as “High” (.25), “Medium” (.15), and “Low” (.05), which covers a range of reasonable values similar to such values in simulation studies for DCMs. These conditions are not meant to be exhaustive. Instead, we set parameter values such that the combinations described game conditions that were both plausible and easily interpretable. We used one reference (i.e., expert) pattern for the initial runs reported here, which came from an analysis of actual game data.

Thus, there are a total of 26 learner trajectories for each SKIVE elements and a total of 21 task parameter conditions arising from seven difficulty conditions crossed with three specificity conditions. This alone results in a total of  $26 \times 21 = 546$  conditions. Since the learning trajectories could be set to equality across different SKIVE elements or to distinct trajectories this number quickly increases in a comprehensive simulation study. Additional factors that we have started to vary in our simulation studies include the nature of the reference / expert pattern, the number of SKIVE elements, and whether learning trajectories are continuous or display piece-wise trends; we will also investigate additional ENA statistics. Thus, the simulation study that we will run during the summer of this year will eventually contain several thousand conditions. For the purpose of this paper we focus predominantly on the core set of 546 conditions because the primary purpose of this paper is to give a flavor of the kind of work that is currently undertaken without claims to comprehensiveness.

### ***Data Generation and Outcome Statistics***

All data generation, compiling, and aggregation was done in *R* ([www.project-R.org](http://www.project-R.org)); the *R* code is available from the first author upon request. In the first step we generated the strings of ‘0’s and ‘1’s that characterize the observable response data by constructing probability matrices for data matrices using equations 6 and 7, then drawing from a Bernoulli distribution for each cell in the data table, and then matching the outcome of the draw with the expert response. For example, if the expert entry for node ‘*I*’ was ‘0’, the probability for this entry was .86, and a ‘1’ was drawn from the Bernoulli distribution representing the event of interest, a ‘0’ was effectively recoded in the data set as that was the outcome of interest. We replicated this process 100 times.

In the second step we computed what we would call ‘*empirical confidence bands*’ for all outcome statistics. For each time slice we computed the lower and upper percentile for the distribution of the ENA statistic of interest – here the weighted density – across all 100 replications. In this study we used the 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles of the distribution to generate 95% empirical confidence bands. In the third step we then computed the *percentage overlap* between the confidence bands for learners with different learning trajectories by counting the number of time slices for which the respective 95% confidence bands overlapped. For example, if the confidence bands overlapped in 62 out of the 87 time slices, this number would have been 71.26%. In general, one would expect this overlap to be higher for learners with similar

underlying trajectories and lower for learners with different underlying trajectories mediated, of course, by the relative magnitudes of the task parameters for the condition of interest.

### Preliminary Results

To obtain an overview of the variation of the percentage overlap statistic for the weighted density across the different learning trajectory pairings, we submitted the values to a factorial ANOVA. With 26 learning trajectory conditions there would have been a total of  $\binom{26}{2} = 325$  different pairs, however. Consequently, the factor ‘type of learning trajectory pair’ would have had 325 levels making comparisons of specific cell means difficult, even with contrasts. Hence, we decided to classify the pairs of trajectories in two different ways.

Specifically, we once coded them using a ‘*surface structure*’ characteristic, which was defined as whether both trends in a pair were linear, curvilinear, or of mixed structure. We did this primarily to illustrate the need for recoding for this paper without expecting this recoding truly to be meaningful because the nature of the trajectory is not really indicative of the similarity of the trends. In other words, a linear and a curvilinear trajectory can, overall, produce very close probabilities of mastery throughout the game play (see, e.g., the slightly curvilinear trajectories and similar linear trajectories) while two linear trajectories can produce starkly different response probabilities (see, e.g., a linear trajectory with zero slope and one with a positive slope). We bring this coding possibility up here only to communicate the methodological rationale, rather than to suggest it for further analyses. Using this recoding scheme we were able to reduce the original set of 325 levels to six levels making meaningful interpretations for a factorial ANOVA much more feasible.

As a more effective coding alternative that gets at the ‘*deep structure*’ of the learning trajectory similarity in a pair, we decided to recode trajectory pairs by whether they produce ‘similar’ probabilities, a term that needed to be operationalized. Learning trajectory pairs were coded as having ‘large degrees of similarity’ if at least 75% of the slices had mastery probabilities for the two trajectories that were within .10 of one another. Similarly, learning trajectory pairs with ‘moderate degrees of similarity’ were defined as having between 25% and 75% of the time slices with probabilities within .10 of one another. Finally, learning trajectories with ‘low degrees of similarity’ were defined as having at most 25% of the time slices with probabilities within .10 of one another. Using this recoding scheme we were able to reduce the

original set of 325 levels to three levels making meaningful interpretations for a factorial ANOVA much more feasible.

The empirical difference of the two different recoding schemes for the factorial ANOVA analyses are shown in Tables 5 and 6 where the effect size for the differences in learning trajectories is markedly higher (i.e., three times the size) for the deep structure recoding than the surface structure recoding. Note that we used a three-way ANOVA model with up to two-way interaction effects – since there is only one weighted density per cell – with the difficulty and specificity conditions for task parameters in Table A1 separated into two factors and the trajectory similarities being the third factor.

Table 5

*Factorial ANOVA Results using Surface Structure Similarity Coding*

<b>Source</b>		<b>Type III</b>		
		<b>SS</b>	<b>df</b>	<b><math>\eta^2</math></b>
Main Effects	Similarity	93.24	5	11.01
	Difficulty	5.22	2	.62
	Specificity	19.22	6	2.27
Interaction Effects	Similarity*Difficulty	.31	10	.04
	Similarity*Specificity	1.35	30	.16
	Difficulty*Specificity	.26	12	.03
Residual		750.11	6759	88.55

Table 5

*Factorial ANOVA Results using Deep Structure Similarity Coding*

<b>Source</b>		<b>Type III</b>		
		<b>SS</b>	<b>df</b>	<b><math>\eta^2</math></b>
Main Effects	Similarity	285.78	2	33.74
	Difficulty	.38	2	.05
	Specificity	1.42	6	.17
Interaction Effects	Similarity*Difficulty	.94	4	.11
	Similarity*Specificity	3.97	12	.47
	Difficulty*Specificity	.26	12	.03
Residual		554.32	6786	65.44

These results therefore suggest that the WD, as a single number, can help differentiate between learners with different underlying trajectories. However, the weighted density statistic is not equally sensitive across different game conditions, which is reflected by the pattern of mean percentage overlap values across different conditions shown in Tables 4 and A2. As we would expect, the weighted density is most sensitive for game conditions in which parameters describing task difficulty,  $1 - s^{(1)}$  and  $g^{(1)}$ , are equally low across all tasks indicating that all tasks are well designed. It is least capable of meaningfully differentiating between learners when those parameters characterize a game in which tasks are poorly designed with both high slipping and high guessing parameters.

Table 4

*Average Percentage Overlap of Weighted Density across Task Parameters Conditions*

	Mean	Median
A	0.65	0.70
B	0.61	0.64
C	0.56	0.56
1	0.52	0.49
2	0.74	0.95
3	0.60	0.62
4	0.62	0.68
5	0.61	0.64
6	0.54	0.53
7	0.59	0.59

As an additional representation of these patterns in the percentage overlap statistic for the weighted density consider Figures A3, A4, and A5, which show the observed values of this statistic for learner pairs with different trajectories for three different task parameter conditions. In these three conditions all tasks are well designed but the primary parameter describing the *task specificity*,  $1 - s^{(0)}$ , varies across the three conditions. In the first task parameter condition (A1) tasks are highly specific. Thus, a learner who has mastered a particular SKIVE element has a low probability of using it unnecessarily when it is not an essential part of the solution strategy for that task. In the second task parameter condition (B1) tasks are moderately specific. Thus, a learner who has mastered a particular SKIVE element has a moderately high probability of using

it unnecessarily when it is not essential for the solution strategy for that task. In the third task parameter condition (C1) tasks are not very specific. Thus, a learner who has mastered a particular SKIVE element has a high probability of using it unnecessarily when it is not essential for the solution strategy for that task. Note that the secondary parameter  $g^{(0)}$  is set to 1 in all three conditions, which is necessary because it is unrealistic to allow learners to demonstrate skills when they are not required and they have not mastered them.

A close inspection of the values for the percentage overlap statistic using 95% confidence bands supports earlier observations. That is, as expected, when the underlying learning trajectories are more similar in terms of the probabilities of mastery for these games, the percentage overlap of the confidence bands for the weighted density is higher and vice versa. This is true within sets of trends (e.g., the percentage overlap of L1\_1 with L1\_3, L2\_3, and L3\_3 under condition A1 is 87%, 59%, and 38% respectively) and across sets of trends (e.g., the percentage overlap of L1\_1 with CL1\_3 and CL2\_3 under condition A1 is 59% and 15%, respectively). The strength of this relationship is mediated by the task parameters, however, such that more constrained tasks make differentiations harder while less constrained tasks make differentiations easier. In other words, weighted density becomes more sensitive to trajectory differences when students are able to use SKIVE elements that are not essential for the solution strategy indicated by the expert.

Overall, these results reflect the expected pattern that the weighted density is driven by the absolute number of occurrences in the cumulative adjacency matrices but quantifies the behavior of the statistic in a more fine-grained manner beyond this general observation. This behavior of the weighted density can be viewed as positive if the use of many SKIVE elements are generally encouraged (i.e., when effective yet relatively inefficient solutions are encouraged). It can also be viewed as negative if efficacious solutions are the target of the game play. Since the latter is more common in epistemic game play the use of the weighted density is not recommended for those contexts and alternative statistics such as the relative centrality or certain distance measures are preferable. These measures are currently investigated in our extended simulation study.

### Ongoing and Future Work

This paper introduces and provides a snapshot of preliminary results from an ongoing research program applying modern psychometric methods to games-based assessment. Over the course of the summer, different ENA statistics, reference / expert patterns, and variations of task conditions will have been evaluated. For example, Figure 4 shows the graphs for piecewise learning trajectories that we are using in one component of the simulation study.

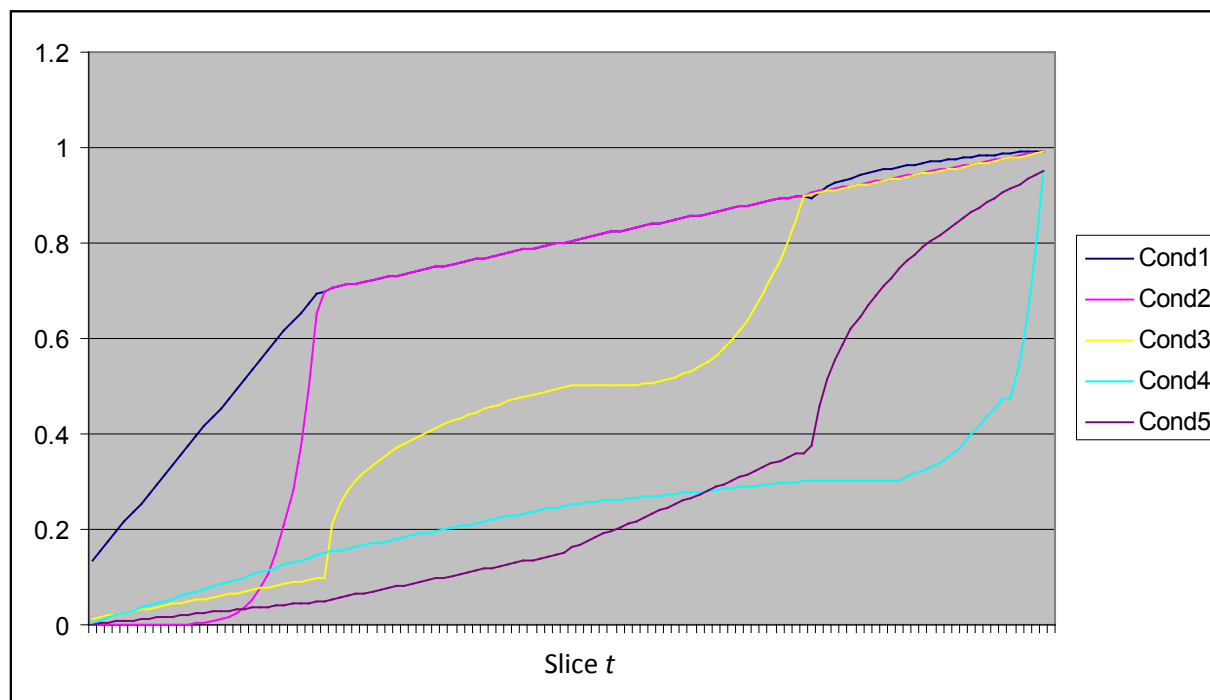


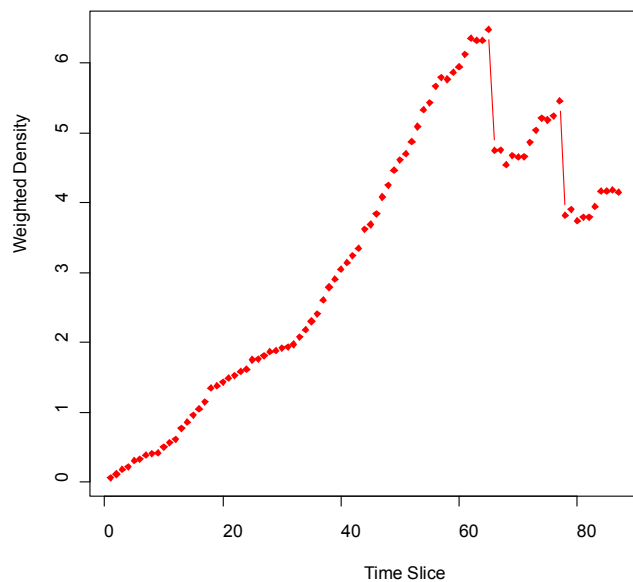
Figure 4 Piecewise learning trajectories for mastery probabilities of learners.

Furthermore, we will vary the learning trajectories across different SKIVE elements, we will vary task parameters across blocks of time slices within a game, and we will implement a random sampling of task parameters. At a conceptual level we are also exploring how different parametric latent-variable models can be applied to subcomponents of epistemic games and how the restructuring of the data for analysis purposes changes the nature of the associated inferences that can be made with different models.

In addition, we are pursuing an investigation of how ENA statistics perform relative to a probabilistic approach for network representation from social network analysis. Specifically, we have used *Pathfinder network analysis* to explore the conditions under which weighted density statistics most accurately reflect reliable connections between SKIVE elements that exist at the

population level. Pathfinder network analysis employs an algorithm to derive minimally noisy networks from observed matrices in which noisy connections between network nodes have been eliminated, thus maintaining only maximally informative connections between nodes (see Schvaneveldt, Durso, & Dearholt, 1989, for a technical description of this method). Pathfinder is described in the literature as a tool for “characteriz[ing] the underlying structure in sets of proximities” (p. 249, Schvaneveldt et al., 1989); it has been used previously to produce concept maps (Schvaneveldt et. al., 1989) and meaningfully differentiate between experts and novices based on individuals’ knowledge organization or cognitive structures (Acton, Johnson, & Goldsmith, 1994; Gomez, Hadfield, & Housner, 1996; Gonzalvo, Cañas, & Bajo, 1994; Schvaneveldt et. al., 1989; Schvaneveldt, Durso, Goldsmith, Breen, Cooke, Tucker, & DeMaio, 1985).

To pursue this line of research we use the same data-generation design as for the core study in this paper and then apply the Pathfinder algorithm to the generated data to see how Pathfinder network representations and associated statistics differ from ENA network representations and ENA statistics. We started these studies by looking at differences between key ENA statistics for cumulative adjacency matrices of the raw data, to which they would normally be applied, and the reduced cumulative adjacency matrices after the Pathfinder algorithm had been applied.



*Figure 5* Average weighted density differences under a linear learning trajectory.

Initial findings from these analyses are consistent with results from the simulation study reported in this paper. For example, Figure 5 shows these differences for the three prototypical task parameter conditions that we described earlier. Contrary to the presentations in the core simulation study reported in this paper, this graph is for learners with specific learning trajectories, however, and not for pairs of learners with different learning trajectories, which is why the weighted density itself and not percentage overlap of weighted density is used.

Figure 5 shows that the differences between the weighted densities for raw and reduced cumulative adjacency matrices are similar across the three task parameter conditions up until about halfway through the game (i.e., that the Pathfinder algorithm consistently eliminates many paths early on during game play when few connections exist in the first place). These differences change later in the game when differences get larger as the tasks become more constrained. This makes intuitive sense because more constrained tasks allow for fewer connections to appear and, thus, more connections being eliminated by the Pathfinder algorithm. When tasks are less constrained, learners can rely on SKIVE elements more frequently, thus creating more occurrences and co-occurrences of SKIVE elements of which fewer get eliminated by the Pathfinder algorithm. Again, though, these patterns match what one would expect given what is known about the Pathfinder algorithm but they quantify the behavior of the weighted density under different conditions in a more fine-tuned manner.

### **Concluding Remarks**

Since ENA is a purely descriptive method, it has several weaknesses vis-à-vis firmly established latent-variable models. At the same time, modern measurement models cannot be easily applied to the kinds of data that arise in epistemic games so searching for viable alternatives and investigating their potential utility is key. One of the main objectives of this paper was to provide readers with a sense of the kinds of methodological considerations and approaches that we have been using in our research program so far. The research program that we have conceptualized so far is specifically designed to explore the utility of alternative modeling approaches for representing the development of individual learners and differences across different learners with different characteristics under different game play conditions for epistemic games. We are at a point in our work where we have clear ideas for where we would like to take the research next, but constructive criticism and productive suggestions are clearly welcome.

We believe that our research program can contribute meaningfully to a critical dialogue about how learning can be best represented in innovative digital learning environments. Our work does not exist in a vacuum, of course. Alternative modeling approaches that are currently used by other researchers are, for example, Bayes nets (e.g., Levy & Mislevy, 2004; Shute et al., 2009), neural networks and hidden Markov models (e.g., Soller & Stevens, 2007), and clustering algorithms (e.g., Nugent, 2009). We note in closing that each approach requires particular design constraints to be met so that the resulting data structures can be properly analyzed with these methods. We also note that work in this area provides new opportunities and challenges for integrating principles from data-mining, traditional multivariate data analysis, and modern measurement. This stems from the fact that fine-grained raw data such as log-files, clickstreams, or chats have to be mined and aggregated first before they can serve as indicators for higher-level analyses, independent of which analytical method is chosen.

Specifically, each approach requires a careful design of the game environment for assessment purposes, which is equally true of the epistemic game environments to which ENA is applied. In other words, by switching from traditional latent-variable methods to ENA we are certainly not able to “analyze away” the core requirement of any assessment enterprise, namely sound assessment design (see Rupp, Gushta, Mislevy, & Shaffer, 2010). We thus want to end on a note that is somewhat of a call to arms for researchers in educational and psychological assessment, namely to explore the utility of different existing and novel modeling approaches for these new learning environments. If we want to transform educational practices with these environments and if measurement specialists want to have a seat at the main table, rather than at the disregarded fringes of such interdisciplinary endeavors, more of this research is clearly needed. As a colleague recently put it, the work in this interdisciplinary area may often feel like “dirty” measurement in many ways – it certainly is by far not as “clean” as the work on traditional large-scale standardized assessments – but the time seems indeed right to get our hands dirty.

### References

- Acton, W. H., Johnson, P. J., and Goldsmith, T. E. (1994). Structural knowledge assessment: Comparison of referent structures. *Journal of Educational Psychology*, 86(2), 303-311.
- 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.
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- 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.
- Gomez, R. L., Hadfield, O. D., and Housner, L. D. (1996). Conceptual maps and simulated teaching episodes as indicators of competence in teaching elementary mathematics. *Journal of Educational Psychology*, 88, 572-585
- Gonzalvo, P., Cañas, J., and Bajo, M. T. (1994). Structural representations in knowledge acquisition. *Journal of Educational Psychology*, 86(4), 601-616.
- Nugent, R. (2009, July). Clustering in cognitive diagnosis: Some recent work and open questions. Presented during the CDMWG meeting at SAMSI 2009, Raleigh, NC.
- 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)
- 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. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford Press.

- Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *Journal of Technology, Learning, and Assessment*, 8(4). Available online at <http://escholarship.bc.edu/jtla/vol8/4>
- Rupp, A. A., Choi, Y., Gushta, M., Mislevy, R. J., Bagley, E., Nash, P., Hatfield, D., Svarowski, G., & Shaffer, D. (2009). 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.
- Schvaneveldt, R. W., Durso, F. T., and Dearholt, D. W. (1989). Network structures in proximity data. In G. Bower (Ed.), *The Psychology of Learning and Motivation: Advances in Research and Theory, Vol. 24* (pp. 249-284). New York: Academic Press.
- Schvaneveldt, R. W., Durso, F. T., Goldsmith, T. E., Breen, T. J., Cooke, N. M., Tucker, R. G., and DeMaio, J. C. (1985). Measuring the structure of expertise. *International Journal of Man-Machine Studies*, 23, 699-728
- 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., 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., Masduki, I., Donmez, O., Dennen, V. P., Kim, Y.-J., Jeong, A. C., & Wang, C.-Y. (2009). *Assessing 21<sup>st</sup> century knowledge and skills in game environments*. Manuscript submitted for publication.
- Soller, A., & Stevens, R. (2007). *Applications of stochastic analyses for collaborative learning and cognitive assessment* (IDA Document D-3421). Arlington, VA: Institute for Defense Analysis.

West, P., Rutstein, D. W., Mislavy, R. J., Liu, J., Levy, R., DiCerbo, K. E., Crawford, A., Choi, Y., & Behrens, J. T. (2009, June). *A Bayes net approach to modeling learning progressions and task performances*. Presented at the Learning Progressions in Science (LeaPS) conference, Iowa City, IO.

Yin, P. and Fan, X. (2001). Estimating  $R^2$  shrinkage in multiple regression: A comparison of different analytical methods. *The Journal of Experimental Education*, 69(2), 203-224.

### **Author Note**

This work 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 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 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. The authors would also like to thank specifically Dr. David W. Shaffer at the University of Wisconsin at Madison, who enthusiastically pushes for quantitative investigations into ENA and inspired us to pursue this work in a serious and principled manner.

## Appendix A

Table A1

Task Parameters of SKIVE Nodes across Difficulty and Specificity Conditions

Difficulty Condition	S		K		I		V		E		Description
	$s^{(1)}$	$g^{(1)}$	$s^{(1)}$	$g^{(1)}$	$s^{(1)}$	$g^{(1)}$	$s^{(1)}$	$g^{(1)}$	$s^{(1)}$	$g^{(1)}$	
1	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	All tasks are well-designed (i.e., all tasks are well designed for slipping and guessing across skills)
2	High	High	High	High	High	High	High	High	High	High	All tasks are poorly designed (i.e., all tasks are poorly designed for slipping and guessing across skills)
3	Low	High	Low	High	Low	High	Low	High	Low	High	All tasks are moderately well designed (i.e., all tasks are well designed for slipping and poorly designed for guessing across skills)
4	High	Low	High	Low	High	Low	High	Low	High	Low	All tasks are moderately well designed (i.e., all tasks are poorly designed for slipping and well designed for guessing across skills)
5	Low	High	Low	High	High	Low	High	Low	High	Low	Tasks are differentially well designed (i.e., tasks are poorly designed for guessing but well designed for slipping on basic skills but well designed for guessing and poorly designed for slipping for complex skills)
6	Low	Low	Low	Low	High	Low	High	Low	High	Low	Tasks are differentially well designed (i.e., tasks are well designed for both guessing and slipping for basic skills while they are well designed for guessing but poorly designed for slipping for complex skills)
7	Low	High	Low	High	Low	Low	Low	Low	Low	Low	Tasks are differentially well designed (i.e., tasks are well designed for slipping but poorly designed for guessing for basic skills but well designed for both guessing and slipping for complex skills)

*(continued)*

Specificity Condition	S		K		I		V		E		Description
	$s^{(0)}$	$g^{(0)}$	$s^{(0)}$	$g^{(0)}$	$s^{(0)}$	$g^{(0)}$	$s^{(0)}$	$g^{(0)}$	$s^{(0)}$	$g^{(0)}$	
A	Low	1	Low	1	Low	1	Low	1	Low	1	All tasks are highly specific (i.e., they provide few opportunities to demonstrate skills that have not been mastered or make successful suppression of skills for an efficient response easy)
B	Medium	1	Medium	1	Medium	1	Medium	1	Medium	1	All tasks are moderately specific (i.e., they provide some opportunities to demonstrate skills that have not been mastered or make successful suppression of skills for an efficient response moderately difficult)
C	High	1	High	1	High	1	High	1	High	1	All tasks are not very specific (i.e., they provide many opportunities to demonstrate skills that have not been mastered or make successful suppression of skills for an efficient response very difficult)

*Note: High = .25, Medium = .15, Low = .05. Conditions A1, B1, and C1 are created by combining the difficulty condition 1 with the specificity conditions A, B, and C.*

Table A2  
*Weighted Density Overlap by Task Conditions*

	<i>N</i>	Mean	Median
<b>A</b>	<b>2275</b>	<b>0.65</b>	<b>0.70</b>
1	325	0.56	0.54
2	325	0.80	1.00
3	325	0.65	0.68
4	325	0.66	0.75
5	325	0.65	0.72
6	325	0.57	0.55
7	325	0.62	0.68
<b>B</b>	<b>2275</b>	<b>0.61</b>	<b>0.64</b>
1	325	0.53	0.49
2	325	0.74	0.94
3	325	0.60	0.59
4	325	0.63	0.69
5	325	0.61	0.64
6	325	0.54	0.52
7	325	0.59	0.60
<b>C</b>	<b>2275</b>	<b>0.56</b>	<b>0.56</b>
1	325	0.49	0.43
2	325	0.69	0.84
3	325	0.56	0.57
4	325	0.58	0.56
5	325	0.57	0.57
6	325	0.51	0.47
7	325	0.55	0.54

	L1 1	L1 2	L1 3	L2 1	L2 2	L2 3	L3 1	L3 2	L3 3	CL1 1	CL1 2	CL1 3	CL2 1	CL2 2	CL2 3	CT 1	CT 2	CT 3	CT 4	CT 5	CT 6	CT 7	CT 8	CT 9	CT 10	CT 11	REF	
L1 1	1		0.67	1	0.89	0.59	1	0.39	0.38	1	0.99	0.59	0.55	0.20	0.15	0.54	0.72	0.90	1	1	1	0.20	0.15	0.16	0.11	0.03	0.02	
L1 2		1		1	1	0.89	1	0.72	0.41	1	0.52	0.49	0.66	0.28	0.15	0.49	0.63	0.74	0.83	1	1	0.64	0.34	0.16	0.11	0.03	0.01	
L1 3			1		1	1	1	1	0.71	1	0.49	0.45	1	0.44	0.15	0.45	0.53	0.69	0.75	0.84	1	0.99	0.55	0.22	0.11	0.03	0.01	
L2 1				1	1	1	1	1	0.82	1	0.44	0.38	0.75	0.52	0.28	0.38	0.52	0.75	1	1	1	0.44	0.20	0.14	0.03	0.02	0.02	
L2 2					1	1	1	1	1	0.78	0.75	0.37	0.94	0.56	0.28	0.37	0.45	0.59	0.72	0.89	1	1	0.90	0.20	0.14	0.03	0.02	
L2 3						1	1	1	1	1	0.40	0.28	0.28	1	0.99	0.40	0.28	0.43	0.52	0.59	0.74	0.89	1	1	0.85	0.15	0.03	0.02
L3 1							1	1	1	0.82	0.44	0.17	0.16	0.94	0.71	0.52	0.20	0.20	0.52	0.72	1	1	1	0.68	0.20	0.11	0.03	
L3 2								1	1	1	0.14	0.14	0.14	1	0.89	0.62	0.17	0.17	0.37	0.56	0.68	0.89	1	1	1.00	0.38	0.14	0.06
L3 3									1	1	0.14	0.14	0.14	1	1	1	0.15	0.15	0.38	0.52	0.59	0.69	0.85	1	1	0.85	0.14	0.03
CL1 1										1	1	0.67	0.41	0.14	0.11	0.66	0.76	0.84	1	0.84	0.53	0.26	0.10	0.10	0.03	0.02	0.01	
CL1 2											1	1	0.28	0.14	0.11	0.83	0.94	1	1.00	0.39	0.22	0.14	0.06	0.10	0.03	0.02	0.01	
CL1 3												1	0.28	0.14	0.11	1	1	1	0.68	0.24	0.15	0.14	0.06	0.10	0.03	0.02	0.01	
CL2 1													1	1	0.33	0.28	0.41	0.49	0.59	0.68	0.77	0.95	1	0.77	0.16	0.03	0.02	
CL2 2														1	1	0.17	0.17	0.28	0.44	0.52	0.59	0.74	0.97	1	1	0.11	0.02	
CL2 3															1	0.11	0.14	0.15	0.18	0.41	0.45	0.56	0.75	1	1	0.99	0.14	
CT 1																1	1	0.89	0.52	0.20	0.18	0.15	0.06	0.10	0.03	0.02	0.01	
CT 2																	1	1	1	0.43	0.28	0.15	0.11	0.14	0.11	0.02	0.01	
CT 3																		1	1	1	0.63	0.20	0.15	0.17	0.11	0.03	0.02	
CT 4																			1	1	1	0.54	0.17	0.17	0.14	0.03	0.02	
CT 5																				1	1	0.84	0.53	0.41	0.15	0.06	0.03	
CT 6																					1	1	0.98	0.51	0.18	0.11	0.03	
CT 7																						1	1	0.74	0.40	0.14	0.07	
CT 8																							1	1	0.75	0.20	0.14	
CT 9																								1	1	0.45	0.16	
CT 10																									1	1	0.44	
CT 11																										1	1	
REF																											1	

All computations are for r = 100 replications and 95% confidence bands

- Overlap within same trend block
- Overlap with trend block of same basic nature (linear, curvilinear, constant)
- Overlap of linear and curvilinear trend blocks
- Overlap of linear, curvilinear, and constant trend blocks

Figure A1 Percent overlap of 95% confidence bands for WD for different learning trajectories under game condition A1.

	L1 1	L1 2	L1 3	L2 1	L2 2	L2 3	L3 1	L3 2	L3 3	CL1 1	CL1 2	CL1 3	CL2 1	CL2 2	CL2 3	CT 1	CT 2	CT 3	CT 4	CT 5	CT 6	CT 7	CT 8	CT 9	CT 10	CT 11	REF			
L1 1	1									1	0.71	0.49	0.51	0.17	0.14	0.49	0.68	0.90	1	1	1	0.16	0.14	0.11	0.03	0.02	0.01			
L1 2		1								1	0.44	0.44	0.66	0.37	0.16	0.44	0.59	0.75	0.87	1	1	0.99	0.24	0.11	0.05	0.03	0.02			
L1 3			1							1	0.41	0.41	1	0.34	0.15	0.41	0.52	0.60	0.72	0.78	0.95	1.00	0.44	0.22	0.03	0.03	0.02			
L2 1				1						1	0.72	0.49	0.87	0.28	0.28	0.86	0.44	0.18	0.25	0.53	0.69	0.95	1	1	0.28	0.17	0.06	0.02		
L2 2					1					1	0.78	0.49	0.51	0.28	0.28	1	0.53	0.28	0.28	0.44	0.59	0.76	0.94	1	1	1.00	0.26	0.06	0.02	
L2 3						1				1	0.97	0.49	0.28	0.28	0.28	1	0.99	0.28	0.25	0.43	0.53	0.70	0.75	0.89	1	1	0.76	0.09	0.06	0.02
L3 1							1			1	0.77	0.49	0.39	0.14	0.15	0.84	0.67	0.49	0.14	0.20	0.44	0.77	1	1	1	0.61	0.25	0.14	0.11	
L3 2								1		1	1	1	0.17	0.11	0.14	1	0.78	0.53	0.11	0.20	0.37	0.59	0.74	0.95	1	1	1.00	0.20	0.14	0.10
L3 3									1	1	1	1	0.13	0.10	0.10	1	1	1	0.10	0.15	0.28	0.45	0.55	0.71	0.87	1	1	0.31	0.16	0.11
CL1 1										1	1.00	0.66	0.20	0.15	0.11	0.66	0.75	0.89	1	0.76	0.41	0.22	0.11	0.03	0.02	0.01	0.01			
CL1 2											1	1	0.20	0.13	0.09	0.80	0.93	1	0.71	0.34	0.14	0.10	0.10	0.03	0.02	0.01	0.01			
CL1 3												1	0.20	0.13	0.09	0.95	1	1	0.48	0.21	0.14	0.10	0.10	0.03	0.02	0.01	0.01			
CL2 1													1	1	0.39	0.17	0.38	0.49	0.60	0.66	0.76	1	1	0.95	0.10	0.03	0.02			
CL2 2														1	1	0.13	0.17	0.28	0.44	0.49	0.60	0.76	0.98	1	0.97	0.08	0.03			
CL2 3															1	0.09	0.11	0.16	0.18	0.28	0.44	0.55	0.75	1	1	0.93	0.43			
CT 1																1	1	0.77	0.39	0.15	0.14	0.10	0.10	0.03	0.02	0.01	0.01			
CT 2																	1	1	0.76	0.59	0.20	0.14	0.14	0.11	0.03	0.03	0.02			
CT 3																		1	1	0.53	0.38	0.15	0.11	0.03	0.03	0.02				
CT 4																			1	1	0.93	0.67	0.28	0.17	0.09	0.06	0.02			
CT 5																				1	1	0.82	0.54	0.28	0.14	0.06	0.03			
CT 6																					1	1	0.91	0.46	0.15	0.13	0.06			
CT 7																						1	1	0.84	0.31	0.14	0.13			
CT 8																							1	1	0.54	0.18	0.15			
CT 9																								1	1	0.43	0.38			
CT 10																									1	1	0.95			
CT 11																										1	1			
REF																											1			
L	Linear learning trajectories (in three blocks L1, L2, and L3 with different intercepts and three slopes '1', '2', and '3' within each)																													
CL	Curvilinear learning trajectories (in two blocks CL1 and CL2 with three different exponents '1', '2', and '3' within each)																													
CT	Constant learning trajectory (11 values in steps of .10 from '0' to '1')																													
	Overlap within same trend block																													
	Overlap with trend block of same basic nature (linear, curvilinear, constant)																													
	Overlap of linear and curvilinear trend blocks																													
	Overlap of linear, curvilinear, and constant trend blocks																													
All computations are for r = 100 replications and 95% confidence bands																														

Figure A4 Percent overlap of 95% confidence bands for WD for different learning trajectories under game condition B1.

	L1_1	L1_2	L1_3	L2_1	L2_2	L2_3	L3_1	L3_2	L3_3	CL1_1	CL1_2	CL1_3	CL2_1	CL2_2	CL2_3	CT_1	CT_2	CT_3	CT_4	CT_5	CT_6	CT_7	CT_8	CT_9	CT_10	CT_11	REF		
L1_1	1																												
L1_2		1																											
L1_3			1																										
L2_1				1																									
L2_2					1																								
L2_3						1																							
L3_1							1																						
L3_2								1																					
L3_3									1																				
CL1_1										1																			
CL1_2											1																		
CL1_3												1																	
CL2_1													1																
CL2_2														1															
CL2_3															1														
CT_1																1													
CT_2																	1												
CT_3																		1											
CT_4																			1										
CT_5																				1									
CT_6																					1								
CT_7																						1							
CT_8																							1						
CT_9																								1					
CT_10																									1				
CT_11																										1			
REF																											1		

	Overlap within same trend block
	Overlap with trend block of same basic nature (linear, curvilinear, constant)
	Overlap of linear and curvilinear trend blocks
	Overlap of linear, curvilinear, and constant trend blocks

computations are for r = 100 replications and 95% confidence bands

Figure A5 Percent overlap of 95% confidence bands for WD for different learning trajectories under game condition C1.

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

In the *Urban Science* game learners assume the role of urban planners in redesigning neighborhoods of 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 collect data via real-life visits to the neighbourhood, they integrate information via a virtual interface that overlays relevant artefacts onto a geographical map of the neighbourhood, and present their results to real-life city council members. During the course of the game, learners work individually and interact with others. This includes other learners as well as mentors that guide them through the game; interaction is conducted either in real-life settings (i.e., meetings) or in virtual settings (e.g., via e-mail or instant messaging).

To provide a sense of the game interface, Figure B1 shows a screenshot of the main page with an email that the learner received from a virtual character. In the email, a specific task is described (i.e., 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. The game consists of a sequence of four broad tasks that ask learners to develop plans for re-zoning different neighbourhoods in Madison, WI. As shown through the links on the right side of Figure B2, the neighbourhoods in this version of the game are *State Street*, *Schenk-Atwood*, *Northside*, and *Madison East*. Each project consists of similar tasks with the ones for the culminating *Madison-East* project whose links are shown in Figure B2 consisting of (a) an issue statement, (b) a summary plan, and (c) an interactive final presentation.

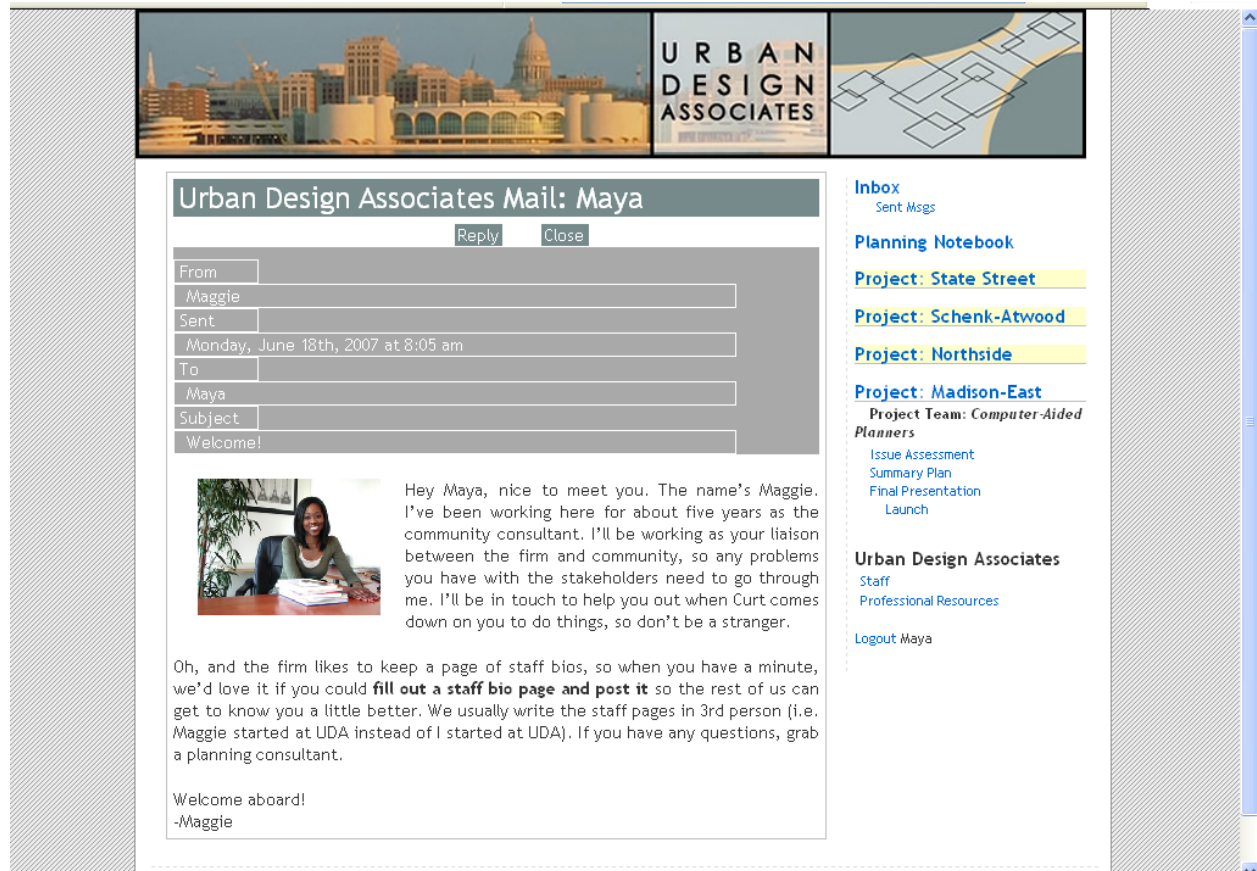


Figure B1 Screenshot of e-mail in the main page of the *Urban Science* interface.

All tasks are similar to one another in that learners are 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. Tasks include, for example, developing a neighbourhood report, developing a preference survey, and developing a presentation that summarizes their proposal. All of these activities have been designed to engage learners in ways that resemble and encourage understanding relevant to the urban planning profession based on ethnographic studies of the urban planning practicum associated with a course at the University of Wisconsin.

As an example, Figure B3 shows part of a final summary proposal for *Madison-East* by one learner group. It shows a map of the neighbourhood along with indicators, whose specific

parameters were developed in collaboration amongst all learners in this group using interactive software designed for creating such maps and indicator representations. 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. This rationale and the choices represent the consensus of the members in that group, who were previously assigned to different stakeholder groups.

Summary Final Plan - Submitted

URBAN DESIGN ASSOCIATES

Inbox  
Sent Msgs

Planning Notebook

Project: State Street

Project: Schenk-Atwood

Project: Northside

Project: Madison-East

Project Team: Computer-Aided Planners

Issue Assessment

Summary Plan

Final Presentation

Launch

Urban Design Associates

Staff

Professional Resources

Logout Maya

Avenue, the main street in the Northside. This allows for increased walkability: Local businesses serve as an attraction by drawing people to the street to shop. This move also increased jobs, meaning that more people who live in the neighborhood can also work and shop there. Above these stores, we zoned high density housing. This housing will allow people to more easily visit local businesses. Lastly we took away the tax incremental funding on major factories to increase property tax revenue. Natural Resources- This plan adds more wetlands to the Northside. This additional wetland can serve as an extra filter

Figure B3 Screenshot of first part of consensual redevelopment plan for *Madison-East*.