Using the ECD Framework to Support Evidentiary Reasoning in the Context of a Simulation Study for Detecting Learner Differences in Epistemic Games

SHAUNA J. SWEET AND ANDRÉ A. RUPP
University of Maryland

The *evidence-centered design* (ECD) framework is a powerful tool that supports careful and critical thinking about the identification and accumulation of evidence in assessment contexts. In this paper, we demonstrate how the ECD framework provides critical support for designing simulation studies to investigate statistical methods within an ill-defined methodological domain like games-based assessment. We discuss the design and selected findings of a complex simulation study to investigate the utility of statistics derived from a non-parametric method called *epistemic network analysis*; this method is used in practice for the analysis of real data from a suite of digital learning environments called *epistemic games*. We present findings that show that one marginal ENA statistic, the *weighted density* statistic from social network analysis, has some ability to detect differences among a variety of simulated learners when they play different types of simulated games. Detection strength ranged from weak to strong, depending jointly on game design characteristics and the types of learners being compared. Our work illustrates the complex challenges of how best to describe, justify, and evaluate design decisions for simulation studies in the context of games-based assessment.

Keywords: Evidence-centered design, simulation study, epistemic games, weighted density, social network analysis

Authors' addresses: Department of Human Development and Quantitative Methodology, University of Maryland, College Park, 20742, shaunajsweet@gmail.com and ruppandr@umd.edu.

1. INTRODUCTION

This special issue is concerned with the design, implementation, and data-analysis for diagnostic assessments in digital learning environments guided by the *evidence-centered design* (ECD) framework [e.g. Mislevy et al. 2006; Mislevy et al. this issue]. In this paper we demonstrate how the ECD framework can also be used to provide critical support for the design of simulation studies that investigate the performance of statistical methods in games-based assessments, which are a relatively ill-defined methodological domain. We specifically report on a simulation study with which we investigated the utility of statistics derived from a non-parametric analytic method called *epistemic network analysis* (ENA) [e.g., Bagley and Shaffer 2010; Nash and Shaffer 2011; Shaffer et al. 2010; see also Rupp et al. 2010], which are applied to the analysis of real data from a suite of digital learning environments called *epistemic games* [e.g. Bagley and Shaffer 2009; Shaffer 2006a; see also www.epistemicgames.org].

1.1 Epistemic Games and Land Science

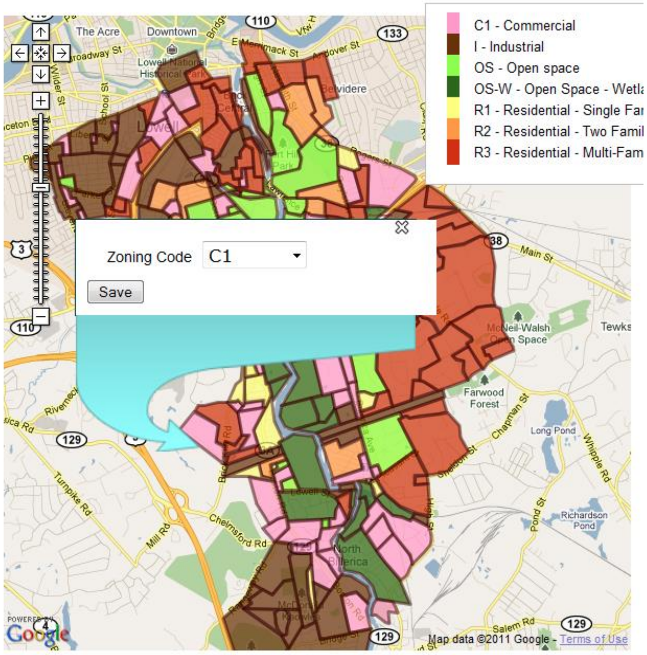
One of the primary educational objectives of epistemic games is the development of ways of reasoning, acting, and communicating that are commensurate with those of professionals in a specialized domain such as urban planning, journalism, or archaeology [Shaffer 2006a]. This could be achieved through an internship in a professional workplace, for example. However, such professional development opportunities may not always be widely available, which is where epistemic games come in.

Epistemic games are designed to help learners develop those skills by engaging them in scenarios that emulate, through a digital medium, the core task objectives, task constraints, and problem-solving processes in which real-life experts in a professional domain engage. We specifically use the example of an epistemic game called *Land Science* in this paper. In *Land Science* learners are presented with a series of authentic complex problem-solving activities around issues of land use, conservation, and community development, in which they engage individually or in groups under the guidance of a *professional mentor*. Learners become interns at the office of a fictitious urban and regional design firm. They weigh the trade-offs of land use decisions in ecologically-sensitive areas, interact with virtual stakeholders, and use interactive tools to develop land use plans for local and national sites. Game play is comprised of approximately four stages, the latter three of which were the ones that we modeled via our simulation study.

At the beginning of the epistemic game learners are welcomed as interns. The first task is the creation of an online biographical web page, which serves as a way for learners to familiarize themselves with the game interface and basic game mechanics. It is also the first time they are exposed to descriptions of the key competencies of domain-specific reasoning that the game targets. The game developers have described these as types of *skills* (S) and *knowledge* (K), facets of *identity* (I), kinds of *values* (V), and ways of *epistemological reasoning* (E) – *SKIVE elements* for short. Theoretically, their joint usage is guided by the *epistemic frame* of the learners [Shaffer 2006b], which is their discipline-specific framework for reasoning, acting, and communicating.

Second, learners engage in a *virtual visit* of the site that is to be re-zoned. Within the context of the game this site visit is much like a *needs analysis*. Learners virtually tour the city area that should be re-zoned. During this tour, learners meet with members of multiple stakeholder groups to which they are randomly assigned (e.g. business community representatives, park and recreation representatives, environmental representatives). Virtual characters from each stakeholder group present learners with their concerns about how land is being used and their ideas for how the land should be rezoned in order to best achieve their goals. During the virtual site visit, contextual and historical information about the city area is made available to the learners as well.

Third, after meeting with other learners from the stakeholder group, learners complete their first proposal, which is called a *preference survey*. This proposal is a simplified version of the game's culminating task. Learners must develop a rezoning proposal that will address the concerns of the stakeholder group they were assigned to. The preference survey is then assessed by the stakeholder group (i.e. via automated scoring rules and human mentors) and feedback is given to learners indicating to what extent they successfully captured that stakeholder group’s objectives. The preference survey and the feedback provided are later used by learners as resources in order to solve the game's final task. The re-zoning activities are done in an interactive interface called *iPlan*, which is shown in Figure 1.



Zoning Code C1

Fig. 1. Screen shot of *iPlan* during re-zoning actions in *Land Science* with a particular parcel being re-zoned using code C1.

Although learners collaborate with one another and discuss how the preference survey is to be completed in light of their stakeholder group‘s interests, each learner submits his or her own proposal and also independently submits justifications for those decisions.

In the fourth stage of the game, learners from all of the different stakeholder groups are mixed and asked to develop a *redevelopment proposal* that integrates the perspectives of all stakeholder groups. This redevelopment proposal is very similar to the preference survey except that learners must now develop a plan which addresses the conflicting needs of multiple stakeholder groups. In other words, for the preference survey learners are assigned to homogeneous groups that represent a single stakeholder group but for the final redevelopment proposal learners are assigned to heterogeneous groups whose responsibility is to represent all different stakeholder groups within the virtual community. The final redevelopment proposal has two components: a final rezoning plan submitted through *iPlan* as well as a notebook entry consisting of a narrative that presents this plan and justifies the proposed redevelopment changes.

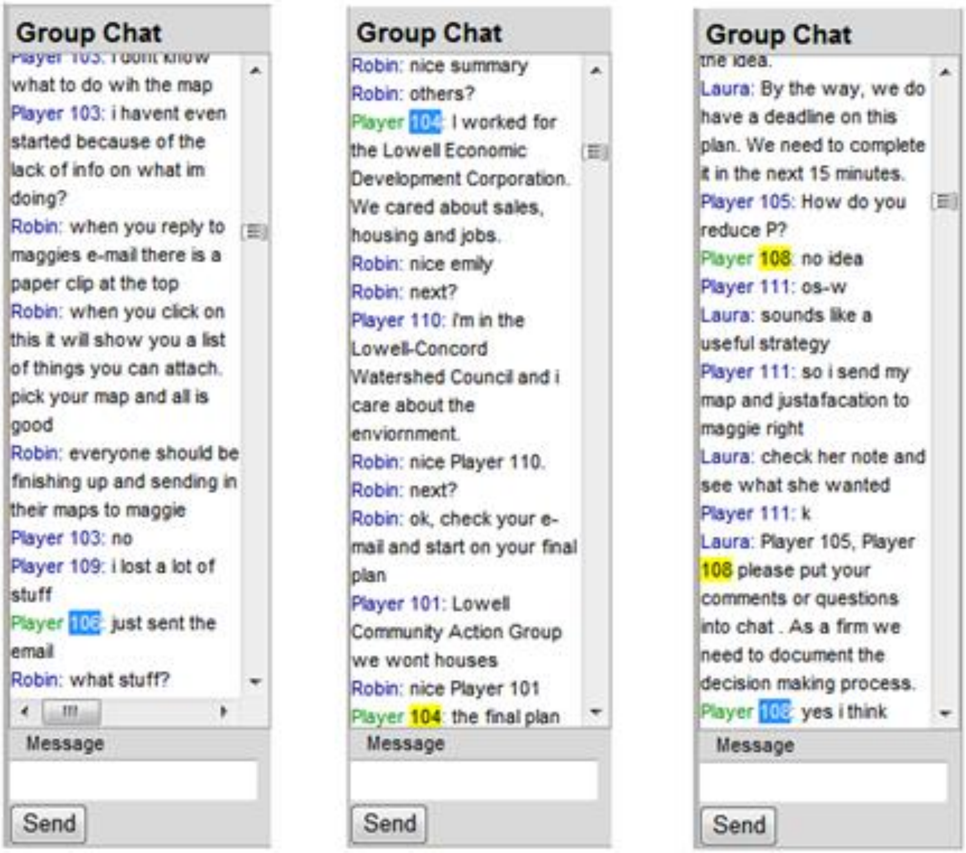
By game design there is generally no single solution that can be construed as optimal in either the homogeneous or heterogeneous stakeholder groups. The key idea of *Land Science* – and urban planning more generally – is to have learners develop consensus-building strategies within both versions of the rezoning task. It is this reasoning through a complex problem that is at the core of epistemic game play: learners have to reconcile diverse stakeholder interests in ways that require reliance on the postulated SKIVE elements as guided by their epistemic frame.

1.2 Evidence Identification and Accumulation in *Land Science*

In ECD terminology, the specification of the SKIVE elements – with a potential breakdown into subelements – as well as a specification of their interrelationships using epistemic frame theory constitutes the specification of the global *student model* for an epistemic game. The individual SKIVE elements are the *student model variables*. Technically, different subsets of SKIVE elements (i.e. student model variables) could be used for different reporting purposes; thus, one could also construct and speak of multiple different student models for different reporting purposes.

Epistemic games like *Land* Science are collaborative, dynamic, and complex, and the process of making inferences about learner development and emergent expertise within epistemic game environments is non-trivial. In ECD parlance, any analysis of the data generated within these environments serves to create an evidence-based narrative. Learners’ observable actions are used to infer their competency profile on the underlying SKIVE elements and to reason about their epistemic frame that guides their use. This reasoning process shares many similarities with the development, implementation, and analysis of complex performance assessments in the domains of educational and psychological measurement [see Mislevy et al. this issue].

In epistemic games like *Land Science*, a key source of observable evidence about SKIVE elements is learners' utterances during online chat as they work together to collaboratively solve problems and complete essential tasks; Figure 2 shows such a chat excerpt.



Player 103: i dont know what to do wih the map
Player 103: i havent even started because of the lack of info on what im doing?
Robin: when you reply to maggies e-mail there is a paper clip at the top
Robin: when you click on this it will show you a list of things you can attach. pick your map and all is good
Robin: everyone should be finishing up and sending in their maps to maggie
Player 103: no
Player 109: i lost a lot of stuff
Player 106: just sent the email
Robin: what stuff?

Robin: nice summary
Robin: others?
Player 104: I worked for the Lowell Economic Development Corporation. We cared about sales, housing and jobs.
Robin: nice emily
Robin: next?
Player 110: i'm in the Lowell-Concord Watershed Council and i care about the enviornment.
Robin: nice Player 110.
Robin: next?
Robin: ok, check your e-mail and start on your final plan
Player 101: Lowell Community Action Group we wont houses
Robin: nice Player 101
Player 104: the final plan

the idea.
Laura: By the way, we do have a deadline on this plan. We need to complete it in the next 15 minutes.
Player 105: How do you reduce P?
Player 108: no idea
Player 111: os-w
Laura: sounds like a useful strategy
Player 111: so i send my map and justafacation to maggie right
Laura: check her note and see what she wanted
Player 111: k
Laura: Player 105, Player 108 please put your comments or questions into chat . As a firm we need to document the decision making process.
Player 108: yes i think

Fig. 2. Chat excerpts for learners playing *Land Science*.

Evidence identification, which is part of the *evidence model* in the *conceptual assessment framework* (CAF) of ECD, involves the segmentation, sorting, and scoring of those utterances. Choices about how this is done are driven by content, sequential, and game design considerations within a particular activity structure. In ECD parlance, how design features of the game form meaningful *evidentiary boundaries* / *evidentiary segments* for evidence identification purposes is part of the *task model* specification.

Figure 3 on the next page illustrates the process of evidence identification in *Land Science*. Learners' utterances are captured during group chat sessions and ordered sequentially in a database. The content of that dialogue is then coded automatically at a fine grain size using a series of *automated scoring rules*. Each scoring rule assigns '0's and '1's to indicate the presence of absence of individual keywords – or particular combinations of keywords – that have been identified by the game developers as indicators that students rely on the underlying SKIVE elements during their activities.

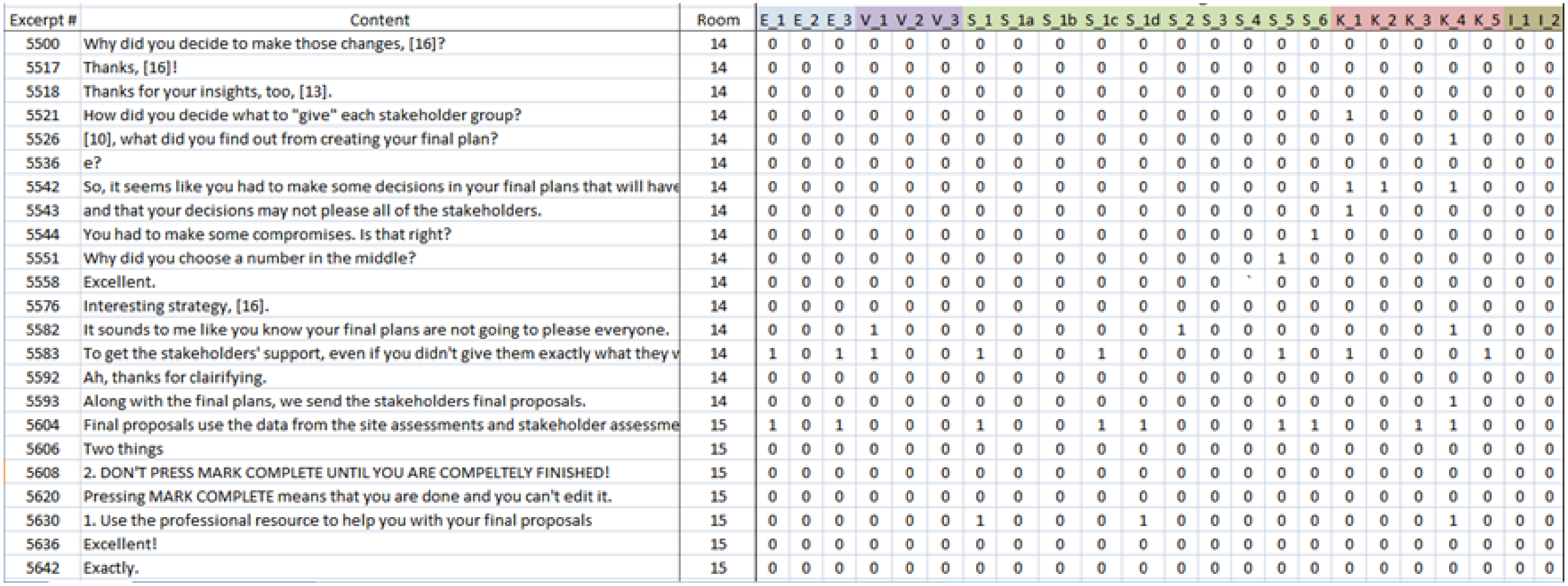
These codes could also be rescored for a more coarse-grained coding for sets of utterances that correspond to meaningful task or thematic boundaries (i.e. learners could receive a '1' if they use relevant keywords in at least one of the constituent utterances within an evidentiary boundary). A similar coding scheme could also be used to aggregate up the binary codes for the 19 subcomponents to create five higher-order SKIVE codes (i.e. learners could receive a '1' for a particular SKIVE element within an evidentiary boundary if they use relevant keywords for at least one of the subcomponents within that boundary).

Once evidentiary boundaries / segments have been identified and utterances have been scored, the resulting vector of binary codes for a particular evidentiary segment and learner is transformed into an *adjacency matrix* using basic logic rules; a sample matrix for a single learner for a single evidentiary segment with five student model variables is shown is shown in Table I. Put simply, an adjacency matrix shows the co-occurrences between SKIVE elements; the entries in the matrix in Table I, for example, indicate that the learner jointly relied on the elements *S*, *K*, and *V* in this particular evidentiary segment.

Table I. Sample Adjacency Matrix for an Individual Learner for a Single Segment / Task

	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

Beliefs about SKIVE elements (i.e. student model variables) get updated across evidentiary segments continuously throughout the epistemic game; in *Land Science* this is most frequently the case when a learner participates in discussions via group chat. The evidence for learners' emerging expertise can then be summarized across any range of evidentiary segments simply by summing the individual entries in the relevant adjacency matrices.



Excerpt #	Content	Room	E_1	E_2	E_3	V_1	V_2	V_3	S_1	S_1a	S_1b	S_1c	S_1d	S_2	S_3	S_4	S_5	S_6	K_1	K_2	K_3	K_4	K_5	I_1	I_2
5500	Why did you decide to make those changes, [16]?	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5517	Thanks, [16]!	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5518	Thanks for your insights, too, [13].	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5521	How did you decide what to "give" each stakeholder group?	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
5526	[10], what did you find out from creating your final plan?	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
5536	e?	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5542	So, it seems like you had to make some decisions in your final plans that will have	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0
5543	and that your decisions may not please all of the stakeholders.	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
5544	You had to make some compromises. Is that right?	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
5551	Why did you choose a number in the middle?	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
5558	Excellent.	14	0	0	0	0	0	0	0	0	0	0	0	0	0	`	0	0	0	0	0	0	0	0	0
5576	Interesting strategy, [16].	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5582	It sounds to me like you know your final plans are not going to please everyone.	14	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
5583	To get the stakeholders' support, even if you didn't give them exactly what they v	14	1	0	1	1	0	0	1	0	0	1	0	0	0	0	1	0	1	0	0	0	1	0	0
5592	Ah, thanks for clairifying.	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5593	Along with the final plans, we send the stakeholders final proposals.	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
5604	Final proposals use the data from the site assessments and stakeholder assessme	15	1	0	1	0	0	0	1	0	0	1	1	0	0	0	1	1	0	0	1	1	0	0	0
5606	Two things	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5608	2. DON'T PRESS MARK COMPLETE UNTIL YOU ARE COMPELTELY FINISHED!	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5620	Pressing MARK COMPLETE means that you are done and you can't edit it.	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5630	1. Use the professional resource to help you with your final proposals	15	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
5636	Excellent!	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5642	Exactly.	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 3. Utterances coded into 19 binary observable variables that are color-coded by the five SKIVE elements.

The resulting matrix is called a *cumulative adjacency matrix*; extending the previous example, Table II shows such a matrix calculated for an individual learner across three evidentiary segments.

Table II. Cumulative Adjacency Matrix for an Individual Learner across Three Evidentiary Segments

	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

This matrix shows, for example, that the learner has used several competencies multiple times, which led to two associations between *S* and *K* as well as *S* and *V*, three associations between *V* and *K*, and one association between the remaining pairs of elements across the segments.

The counts in the cumulative adjacency matrix are viewed by the game developers as evidence of the mastery of the individual competencies as well as the connections between them, the latter being the conceptual hallmark of emerging expertise under epistemic frame theory. In the language of ECD, the creation of cumulative adjacency matrices and the computation of statistics for them is the process of *evidence accumulation*, which is also a part of the *evidence model* in the CAF.

1.3 The Weighted Density Statistic of Epistemic Network Analysis

The data that arise from the above evidence identification and accumulation processes are multivariate in nature, consist of discrete indicators, are collected longitudinally within a game, and may be available only for a small set of learners within a single classroom or program. Consequently, there is currently no prototypical statistical analytic method – certainly no parametric one that we know of – that can be applied directly to these data.

This was the impetus for the development of ENA, which is a non-parametric analysis approach. While different variants of ENA have been proposed in recent years by members of the game development team [e.g. Bagley and Shaffer 2010 use multidimensional scaling in the process and Nash and Shaffer 2011 use tools from social-network analysis] we focus on the social-network based variant of ENA.

We specifically focus on one of the key marginal statistics under this approach, which is known as the *weighted density* (WD); it is defined as follows:

$$WD_t = \sqrt{\sum_{k=1}^{K} \sum_{k \leq k'} a_{kk',t}^2} \qquad (1)$$

where $k = 1,\ldots,K$ indexes a particular SKIVE element and $a_{kk',t}$ is simply a particular entry in the cumulative adjacency matrix at evidentiary segment t (see Table II); only entries in the lower off-triangle are effectively used.

Put differently, the WD statistic summarizes, for each learner, the total number of unique pair-wise associations / connections between SKIVE elements. This summary can be done for a single evidentiary segment, if computed only for the adjacency matrix of that segment, or across multiple evidentiary segments, if computed for the cumulative adjacency matrix across the relevant segments.

1.4 Study Objective & Research Questions

Digital learning environments are becoming increasingly flexible and dynamic while researchers are just beginning to map out the specific learning processes within these environments. While the designs of these environments are often in flux in early stages of development, there is typically an immediate need for understanding how reliable and valid characterizations of learners can be constructed. Ideally, statistics that support such characterizations would be useful for guiding individual learners' interactions with each other and with the system. They might also be used to provide game developers with information about how to fine-tune the design of an epistemic game to optimize learning overall.

The strongest evidence for the performance of any statistic under novel conditions is generally collected via *simulation studies*. Simulation studies are *designed statistical experiments* that can provide reliable scientific evidence about the performance of statistical methods. As noted concisely by Cook and Teo [2011]:

> In evaluating methodologies, simulation studies: (i) provide a cost-effective way to quantify potential performance for a large range of scenarios, spanning different combinations of sample sizes and underlying parameters, (ii) allow average performance to be estimated under repeat Monte Carlo sampling and (iii) facilitate comparison of estimates against the "true" system underlying the simulations, none of which is really achievable via genuine applications, as gratifying as those are. (p. 1)

In our collaboration with the game development team, we sought to investigate the performance of the WD statistic under a range of data-analytic conditions. We specifically wanted to address the following two research questions:

1. How useful is the WD statistic for characterizing the performance of individual learners when they play differently designed epistemic games?

2. How useful is the WD statistic for characterizing the performance of pairs of learners when they play differently designed epistemic games?

In the next section, we describe the design and implementation of the simulation study while in the subsequent section we describe the results of the simulation study. We close this paper with a critical review of the key findings, a discussion of the generalizability of our findings, and a reflection on the use of the ECD framework for simulation study designs in games-based assessment contexts.

2. DESIGNING AND IMPLEMENTING THE SIMULATION STUDY

We designed and implemented this simulation study during early stages of game development so that the results of the study could inform decisions about the game design. Consequently, only some of the design features of the game were stable enough to be operationalized as simulation design conditions because the game developers were still exploring how best to apply what they had learned in their *domain analysis* for urban planning practices to the specification of the learning environment.

This illustrates nicely a tension that many researchers who are working in games-based assessment and related design-based research contexts are facing. On the one hand, design teams want to incorporate some evidence about the performance of their analytic methods early on in the game development process. On the other, statisticians require clear guidelines for operationalizations of game design features to design and implement well-targeted simulation studies that can produce evidence that is both statistically sound and practically relevant.

Negotiating this tension can be particularly challenging within interdisciplinary teams comprised of members who come from different methodological traditions and are accustomed to discipline-specific standards for evidence. We found that the ECD framework supported the design of our simulation study by providing a much-needed conceptual and rhetorical structure that team members could use to effectively communicate with one another. In particular, it helped all members to articulate the core drivers of their targeted evidentiary assessment argument: the characteristics of the learners, the tasks, and their associated interactions.

2.1 Design Summary

Our simulation study design can be viewed as a setup where 21 distinct types of learners – or $\binom{21}{2} = 210$ pairs of learners – play 2 (game length) × 21 (task parameter) × 4 (design matrix) = 168 different games. Table III shows an overview of these design conditions; note that we use the term *task* synonymously with evidentiary segment for the rest of this paper as it lends itself naturally to communicating the evidentiary and conceptual story of this study.

Table III. Summary of Simulation Design Conditions

ECD Model	**Factor**		**Levels**	**# of levels**
Learner Characteristics				
Student Model	# of Student Model Variables		S, K, I, V, E	1
	Type of trajectory		Linear	15
			Curvilinear	6
Total # of learner conditions				21
Game characteristics				
Task Model	Task specification	Task difficulty	Various types	7
		Task specificity	Low, moderate, high	3
Assembly Model	# of segments / tasks		60, 120	2
	Design matrix specification	Task complexity	Constant, varying	2
		Task content	Constant, varying	2
Total # of game conditions				168
# of replications per learner-by-game condition				100

As we discuss in the next section, we relied on the ECD framework to ensure a defensible mapping of our simulation parameters onto features of the real-life epistemic games. We utilized principles from *item response theory* (IRT) [e.g. de Ayala 2009; Reckase 2009] and *diagnostic classification modeling* (DCM) [e.g. Rupp et al. 2010] to define learner and task parameters. We then characterized the expected behavior of learners with different learning trajectories for games comprised of tasks of varying difficulty, complexity, and content. These characterizations were operationalized as probabilities of successful activation or suppression of SKIVE elements (i.e. student model variables). These probabilities were then used to generate response data so that the WD statistic could be repeatedly computed on the generated data sets across replications.

2.2 Learner Parameters

As noted earlier in the paper, specifying the number of SKIVE elements, their grain size, and their development over time constitutes the core of the specification of the student model. Learners' differential engagement with the game and consequent differential development of expertise was evident in qualitative observations of learners in real-life implementations of different epistemic games as well as from pre / post interviews. Through a series of conversations with the design team about how learners of varying ability levels differentially acquire expertise over the course of game play, we identified several types of learners which we sought to represent in our simulation study.

We first conceived of each of the SKIVE elements (i.e. student model variables) as possessing two *mastery states* ('mastered' or 'non-mastered') at each evidentiary segment / task in the game. We then specified distinct trajectories of mastery probabilities for each of these variables across the different segments / tasks that comprise the game. Some learners were specified to progress toward mastery steadily throughout the game play, some learners were specified to learn quickly at the outset of a game with minimal gains later on, some learners were specified to learn slowly initially with rapid gains later in the game, and some learners were specified as remaining at their initial mastery levels throughout the game.

These learning trajectories were operationalized using 15 linear and six curvilinear trajectories that are shown in Figures 4 and 5, respectively, on the next page. In each of the figures, a point on the X-axis corresponds to an evidentiary segment / task and a point on the Y-axis corresponds to the mastery probability for a particular SKIVE element (i.e. student model variable). Although trajectories were designed to represent different types of learners observed during pilot administrations of the game, they were necessarily an abstraction of how learners develop their epistemic frames over time in real life. To keep the design of the simulation study further manageable – and because more fine-tuned theories of development were not available at the time of the study – we generated data such that the mastery trajectories for all SKIVE elements (i.e. student model variables) were identical.

As noted in Table III above, the results in this paper focus only on conditions with five student model variables representing a relatively coarse definitional grain size of the SKIVE variables. Even though this is computationally realistic, it is also a limitation of this study. Yet, it facilitates simpler descriptions of the resulting patterns as well as a sharper focus on the use of the ECD framework during the study design process for the purposes of this paper.

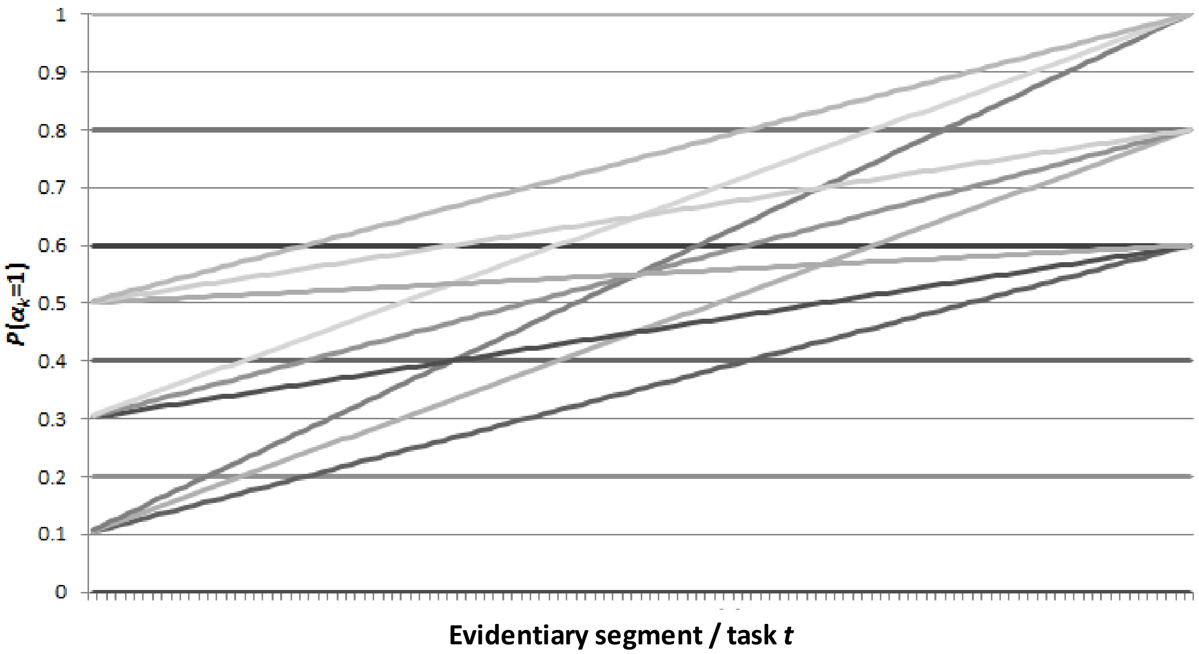


Fig. 4. Linear trajectories of mastery probabilities for SKIVE elements (i.e. student model variables) across evidentiary segments / tasks.

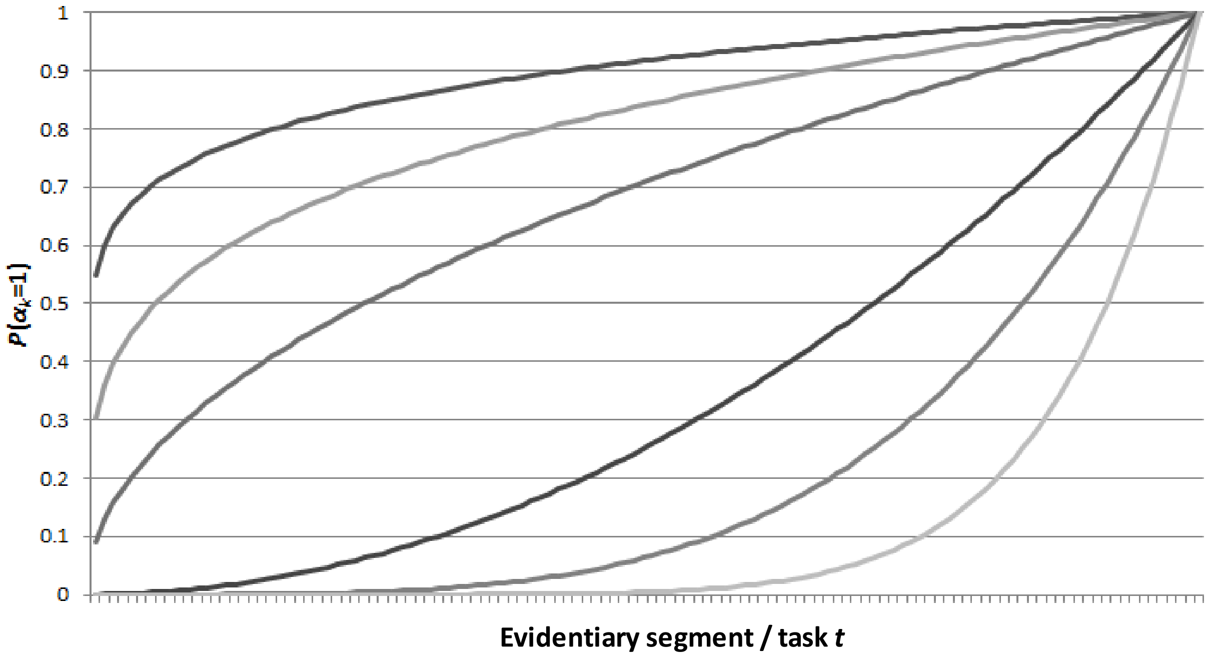


Fig. 5. Curvilinear trajectories of mastery probabilities for SKIVE elements (i.e. student model variables) across evidentiary segments / tasks.

2.3 Specification of Task Parameters

With the game still in development, we relied on the principles of modular task model specification / task design as described within the ECD framework. We operationalized the influence of key task design features via task parameters, which we viewed as reasonably representative abstractions of possible measurement properties.

More concretely, modern latent-variable models for assessment data such as multidimensional IRT models [e.g. Reckase, 2009] contain ‘difficulty’, ‘discrimination’, and ‘guessing’ parameters to reflect the operating characteristics of assessment tasks. DCMs [Rupp et al. 2010; Rupp et al. this issue], which are also multidimensional models, contain analogous task parameters that are formulated relative to the mastery states of the learner; they are referred to as ‘slipping’ and ‘guessing’ parameters in the literature. We chose to use DCMs for the purposes of this study due to slightly simpler operationalizations that result from this choice.

Slipping parameters, typically denoted by the letter s, represent probabilities of responding inappropriately when learners have mastered a particular targeted skill or a set of targeted 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 targeted skill or a set of targeted skills. In other words, $(1 - s)$ and g are the task-specific probabilities that a learner will demonstrate their mastery of a particular student model variable for a given task when it is expected by game design.

Within the context of epistemic games, the focus is not simply whether or not a learner produces a correct response to a particular task, however, either in absolute or in graded terms. Because the pedagogical goal is for learners to develop ways of reasoning through authentic problems with the kind of economic application of skills that characterizes professional expertise, learners are expected to produce *efficacious solutions* to tasks. 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 the key epistemic frame elements that are necessary and sufficient).

To acknowledge this characteristic of epistemic games, we defined four core task parameters for this simulation study that are derivatives of DCM parameters; one obtains eight probabilities from these four because of the complementary events with total probabilities that sum to one for each pair. These parameters describe the extent to which learners’ reliance on SKIVE elements matches the expected activation pattern (i.e. the usage pattern of an expert / professional in the field). We distinguish between guessing and slipping parameters for each SKIVE element (i.e. student model variable) when an expert would utilize (i.e. activate) a particular skill to complete a task (Q_{tk}= 1) and when the expert would not utilize (i.e. suppress) a particular skill to complete a task (Q_{tk} = 0).

We use α to denote one of the K SKIVE elements (i.e. student model variables), Q_{tk} to denote an expected response to task t for SKIVE element k, X_{tk} to denote the observed learner response to task t for SKIVE element k, and a parenthetical superscript to denote whether the slipping or guessing parameter is for the response of '1' (activation) or the response of '0' (suppression). Thus, we defined our eight conditional response probabilities based on our four task parameters as follows:

$$P(X_{tk} = 1 | \alpha_k = 1, Q_{tk} = 1) = 1 - s_{tk}^{(1)} \qquad (2)$$

$$P(X_{tk} = 0 | \alpha_k = 1, Q_{tk} = 1) = s_{tk}^{(1)} \qquad (3)$$

$$P(X_{tk} = 0 | \alpha_k = 1, Q_{tk} = 0) = 1 - s_{tk}^{(0)} \qquad (4)$$

$$P(X_{tk} = 1 | \alpha_k = 1, Q_{tk} = 0) = s_{tk}^{(0)} \qquad (5)$$

$$P(X_{tk} = 1 | \alpha_k = 0, Q_{tk} = 1) = g_{tk}^{(1)} \qquad (6)$$

$$P(X_{tk} = 0 | \alpha_k = 0, Q_{tk} = 1) = 1 - g_{tk}^{(1)} \qquad (7)$$

$$P(X_{tk} = 0 | \alpha_k = 0, Q_{tk} = 0) = g_{tk}^{(0)} \qquad (8)$$

$$P(X_{tk} = 1 | \alpha_k = 0, Q_{tk} = 0) = 1 - g_{tk}^{(0)} \qquad (9)$$

For example, $1 - s_{tk}^{(1)}$ in equation (2) is the probability that a learner who has mastered a particular SKIVE element successfully activates it when expected while $g_{tk}^{(1)}$in equation (6) is the probability that a learner who has not mastered it successfully activates it when expected. Similarly, $1 - s_{tk}^{(0)}$ in equation (4) is the probability that a learner who has mastered a particular SKIVE element successfully suppresses it when expected while $g_{tk}^{(0)}$in equation (6) is the probability that a learner who has not mastered it successfully suppresses it. The remaining four probabilities (3), (7), (5) and (9) are the complements of these.

These slipping and guessing parameters comprise the statistical and conceptual building blocks for the definition of task features according to what we called *task difficulty* and *task specificity*. We defined task difficulty as a function of the slipping and guessing parameter values for the successful activation of SKIVE elements (i.e. $1 - s_{tk}^{(1)}$ and $g_{tk}^{(1)}$) and task specificity as a function of the slipping and guessing parameter values for the successful suppression of SKIVE elements (i.e. $1 - s_{tk}^{(0)}$ and $g_{tk}^{(0)}$).

Again, because a simulation study narrative is a statistical narrative at its core, it was important that decisions about game segmentation and the decision of task parameters reflected efforts to balance statistical and practical considerations. Task parameters needed to be selected in a way that was reflective of an understanding of how the WD statistic is calculated to ensure that we induced sufficient variation in conditions to observe variation in the statistic.

Since it is obviously easy to suppress (i.e. not use) a particular SKIVE element when it has not been mastered, we decided to set $g_{tk}^{(0)} = 1$ for all conditions. For the other three task parameters, "low", "medium", and "high" levels were operationalized by sampling, for each evidentiary segment / task of the game, slipping and guessing parameters for each SKIVE element (i.e. student model variable) from four-parameter Beta distributions with the following interval boundaries and means: *High:* (.20, .30), mean = .25; *Medium:* (.10, .20), mean = .15; *Low:* (.00, .10), mean = .05. The crossing of seven task difficulty conditions (1-7; see Table A-I) with three task specificity conditions (A, B, C; see Table A-II) yielded 21 different task conditions (A1, A2, …,C6, C7).

2.4 Task Assembly / Game Design

Finally, we considered how tasks of varying difficulty and specificity might be combined to create an overall game, the composition of which we captured in four different *design matrices*. Design matrices reflect the way in which multiple tasks are strung together to comprise a complete game that has a plausible structure and flow, and are conceptually tied to what is referred to as the *assembly model* of the game in the ECD framework.

Design matrices are also referred to as *Q-matrices* [e.g. Tatsuoka, 2009; Rupp et al. 2010] in the educational and psychological measurement literature. They are conceptually related to simple *tables of specification* for test designs as they specify which tasks are designed to measure (i.e. provide evidence of mastery of) a (latent) proficiency / dimension. Each evidentiary segment / task corresponds to a row in a design matrix and each of the SKIVE elements (i.e. student model variables) correspond to a column. If a task targets a particular SKIVE element this is indicated by a '1' in the corresponding cell of the design matrix; otherwise a '0' is recorded. Design matrix 1, which we will frequently refer back to later in the paper when we discuss results, is shown in Figure 6 on the next page.

Operationalizing learners' development of mastery over the course of game play requires that meaningful evidentiary segments be defined. As noted earlier, segments may be defined by game content, specifications of task objectives, or dynamically based on learners' interactions with each other. To align with the number of evidentiary segments we had observed in practice, we specified two game lengths of $T = 60$ and $T = 120$ segments / tasks. They were then re-conceptualized as three distinct *blocks of tasks* representing the beginning, middle, and end portions of the activity flow of the game. These decisions were consistent with the structure of the piloted version of *Land Science* which featured three delineable sets of activities which varied in focus and complexity (i.e. virtual site visit, preference survey, and redevelopment proposal).

$$\left[\begin{array}{ccccc} S & K & I & V & E \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \end{array}\right]$$

1-40

$$\left[\begin{array}{ccccc} S & K & I & V & E \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{array}\right]$$

41-80

$$\left[\begin{array}{ccccc} S & K & I & V & E \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{array}\right]$$

81-120

Fig. 6. Design matrix 1, which includes tasks that increase in complexity and focus on different SKIVE elements over the duration of game play.

We then differentiated which SKIVE elements were targeted in each of the three blocks of tasks. For example, in design matrix 1 above the more "foundational" SKIVE elements *S* and *K* were specified to be targeted earlier in the game while more "advanced" SKIVE elements *I*, *V*, and *E* were specified to be targeted later in the game. Table IV summarizes the measurement properties of the four design matrices that we used in our simulation study.

A design matrix can be viewed as a structure wherein two prototypical task complexity designs are crossed with two prototypical SKIVE focus designs. For example, while both design matrices 1 and 2 have a progression from low to medium to high complexity across the three blocks of 40 tasks each, design matrix 2 contains tasks that require the activation of all five SKIVE elements throughout the game whereas design matrix 1 focuses on the more "foundational" competencies *S* and *K* first, then on all SKIVE elements, and at the end specifically on the more "advanced" competencies *I*, *V*, and *E*. Similarly, while the mean and median counts of SKIVE elements across all tasks and the SKIVE focus across the game blocks are identical for design matrix pairs 1 and 3 as well as 2 and 4, they differ in task complexity across the game blocks.

Table IV. Structure of Design Matrices across Three Blocks of 40 Tasks

Q-matrix	**Task Complexity in Game Blocks**			**SKIVE Focus in Game Blocks**			**Marginal Counts across Game Blocks**					**Marginal Means**	
	1	**2**	**3**	**1**	**2**	**3**	**S**	**K**	**I**	**V**	**E**	**Mean**	**Median**
1	Low	Med	High	S, K	All	I, V, E	77	82	72	73	71	3.13	3
2	Low	Med	High	All	All	All	69	76	57	57	47	2.55	2
3	Low	Low	Low	S, K	All	I, V, E	70	77	76	70	83	3.13	3
4	Low	Low	Low	All	All	All	65	70	60	59	52	2.55	2

Note. The actual probability values for the occurrence of SKIVE elements in the three blocks of tasks for the four design matrices are available from the first author upon request.

We note that the assignment of targeted SKIVE elements to particular tasks within the three blocks of tasks in a particular design matrix was done probabilistically so that the relative emphasis on particular SKIVE elements was not absolute. Although we empirically investigated the impact of game length during our analyses, the results in this paper focus only on games comprised of $T = 120$ segments. Results from preliminary analyses showed that learner differences were captured more sharply here versus games comprised of fewer segments.

2.5 Data Generation

Recall that inverse slipping and guessing parameters were defined above as probabilities of successfully activating or suppressing particular SKIVE elements (i.e. student model variables) when completing a particular task conditional on learners' proficiency / mastery state for each variable. In order to generate data we computed, for each learner type, task type, and game type, the probabilities of producing responses of '1' and '0' across both possible mastery states of the SKIVE variables.

As shown in equations (10) – (13), these probabilities are functions of (a) the mastery probabilities for each of the student model variables – $P(\alpha_k = 1)$ for 'mastery' and $P(\alpha_k = 0)$ for 'non-mastery' – defined at each task of the game via the learners' trajectories (see Figures 4 and 5), (b) the slipping and guessing parameters of the tasks (see Tables A-I and A-II), and (c) the design matrix entries Q_{tk}:

$$P(X_{tk} = 1|Q_{tk} = 1) = P(X_{tk} = 1|\alpha_k = 1, Q_{tk} = 1)P(\alpha_k = 1) + P(X_{tk} = 1|\alpha_k = 0, Q_{tk} = 1)P(\alpha_k = 0)$$
$$= (1 - s_{tk}^{(1)})P(\alpha_k = 1) + g_{tk}^{(1)}P(\alpha_k = 0) \qquad (10)$$

$$P(X_{tk} = 0|Q_{tk} = 1) = P(X_{tk} = 0|\alpha_k = 1, Q_{tk} = 1)P(\alpha_k = 1) + P(X_{tk} = 0|\alpha_k = 0, Q_{tk} = 1)P(\alpha_k = 0)$$
$$= s_{tk}^{(1)}P(\alpha_k = 1) + (1 - g_{tk}^{(1)})P(\alpha_k = 0) \qquad (11)$$

$$P(X_{tk} = 0|Q_{tk} = 0) = P(X_{tk} = 0|\alpha_k = 1, Q_{tk} = 0)P(\alpha_k = 1) + P(X_{tk} = 0|\alpha_k = 0, Q_{tk} = 0)P(\alpha_k = 0)$$
$$= (1 - s_{tk}^{(0)})P(\alpha_k = 1) + g_{tk}^{(0)}P(\alpha_k = 0)$$
$$= (1 - s_{tk}^{(0)})P(\alpha_k = 1) + P(\alpha_k = 0) \qquad (12)$$

$$P(X_{tk} = 1|Q_{tk} = 0) = P(X_{tk} = 1|\alpha_k = 1, Q_{tk} = 0)P(\alpha_k = 1) + P(X_{tk} = 1|\alpha_k = 0, Q_{tk} = 0)P(\alpha_k = 0)$$
$$= s_{tk}^{(0)}P(\alpha_k = 1) + (1 - g_{tk}^{(0)})P(\alpha_k = 0)$$
$$= s_{tk}^{(0)}P(\alpha_k = 1) \qquad (13)$$

Recall that we set $g_{tk}^{(0)} = 1$, which implied 1 - $g_{tk}^{(0)} = 0$, so that equations (12) and (13) simplified.

Put differently, each combination of SKIVE mastery probabilities, task parameters, and design matrix entries across evidentiary segments / tasks thus created a probability matrix for each simulated learner who played each type of simulated game. We then performed 100 replications by taking 100 independent Bernoulli draws from these probability matrices and transformed the random draws into an observable dichotomous score variable. This led to 100 generated data sets for each game design condition for each of the 21 learner types and 100 × 100 = 10,000 crossed data sets for pair-wise comparisons of learners. We performed all computations in *R* [R Development Core Team 2008] with code that is available from the first author upon request.

2.5.1 Computation of Two Variants / Aggregations of the WD Statistic. The WD statistic can be calculated at any point during game play to summarize observed accumulated evidence for each learner's emerging expertise as captured by the SKIVE elements (i.e. student model variables). In line with the interests of the game development team at the time, we chose to focus on two different analytic variants /

aggregations of the WD statistic. Both variants could technically be computed at any segment / task of the game, but we chose their computation at the final segment / task for illustrative purposes and simplicity. On the one hand, we examined the performance of the WD statistic when calculated for individual learners once at the conclusion of the game; this led to 100 values for each learner and game design condition based on the 100 replications.

On the other hand, we focused on an aggregate value of the WD statistic for learner pairs, which was calculated across all segments / tasks based on the similarity of WD values at each segment / task. This variant was computed in three steps. First, we computed the WD statistic for the cumulative adjacency matrix at each segment / task for each learner type and game design condition, leading to 100 values at each segment / task. We then computed the 2.5th and 97.5th percentiles of these distributions of WD values at each segment / task for each learner type and game design condition, which resulted in what we would call *empirical pseudo-confidence intervals*. Finally, we computed the percentage of segments / tasks for which the 95% pseudo-confidence bands for a pair of learners overlapped (i.e. for which there was no notably sharp difference between pairs of learners). This resulted in a single percentage-overlap value for each pair of learners based on the 100 replications.

We also performed factorial *analyses-of-variance* (ANOVA) on the resulting two variants / aggregations of the WD statistic with the game design factors as independent variables. The performance of the two WD variants is presented in the following section in three subsections for each, (1) general patterns, (2) ANOVA results, and (3) follow-up graphics for explanatory purposes. Additional results for these two WD variants and other ENA statistics that are not the focus of this paper are available from the first author upon request.

3. Results

3.1 WD Statistic for Individual Learners (Variant 1)

3.1.1 General Patterns. Figure 7 shows the values of the WD statistic for the cumulative adjacency matrix of all 21 learner types at each segment / task over the duration of game play for a game built with design matrix 1. Each line depicts the median WD value calculated across all 100 replications; to aid in the visual interpretability of the patterns, the WD values for six representative learner types with relatively distinct trajectories are highlighted in color.

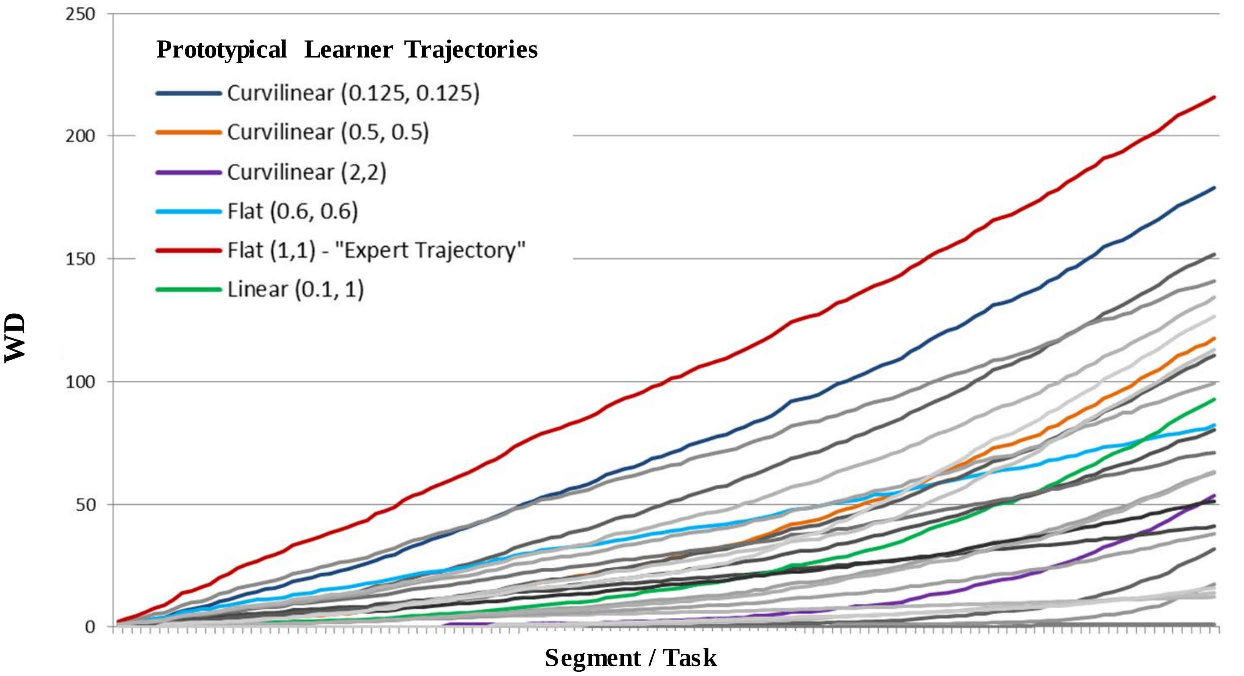


Fig. 7. Median values of the WD statistic for 21 learners for a game designed with design matrix 1.

The meanings of the six highlighted trajectories are as follows:

(a) red corresponds to an expert (*flat trajectory / linear trajectory type 1*)

(b) green corresponds to a type of learner who starts off with low mastery and makes steady gains throughout the game until full mastery is achieved (*linear trajectory type 2*)

(c) light blue corresponds to a type of learner who starts off with moderate mastery and does not progress at all (*linear trajectory type 3*)

(d) dark blue corresponds to a type of learner who makes very quick learning gains early in the game and then makes progressively fewer gains (*curvilinear trajectory type 1*)

(e) orange corresponds to a type of learner who makes quick learning gains early in the game but slower than the learner type shown in blue (*curvilinear trajectory type 2*)

(f) purple corresponds to a type of learner who makes very slow gains initially until a learning spurt at the end of the game (*curvilinear trajectory type 3*)

Figure 7 clearly shows that the trajectory of an expert learner, which is someone with perfect mastery of all SKIVE elements throughout the game, shows the largest values of the WD statistic throughout the game as one would expect. Similarly, one can see that "faster" learner types display WD trajectories that are closer to the expert trajectory, as one would also expect. Faster learners are those who approach high mastery probabilities earlier in the game and, thus, are more likely to utilize the SKIVE elements that are targeted by the tasks for a larger number of tasks compared to “slower” learners. Figure 3 also shows that the WD is more useful for differentiating between learner types when games are played longer because relative differences become more pronounced over time (i.e. more evidence about learner differences can be accumulated).

3.1.2 ANOVA Results. To describe which game design factors are associated with the observed variation in the WD statistic for individual learners, we first conducted 21 four-way ANOVAs, one for each learner type. These analyses included task difficulty (7 conditions), task specificity (3 levels), task content (2 levels), and task complexity (2 conditions) as predictor variables for a total of 84 cells or design conditions. Table IV shows the median importance rankings and associated median effect sizes (i.e. median η^2 values) for the main effects and two-way interaction effects computed across the 21 four-way ANOVAs; the values of the median effect sizes for higher-order interaction effects are not shown here since they are even lower than the ones for the two-way interaction effects.

Table IV. Median η^2 for Four-way Factorial ANOVA for Version 1 of WD Statistic

Effect	Design Factor	Median Rank	Median η^2
Main Effect	Task Complexity	1	.2827
	Task Difficulty	2	.2784
	Task Specificity	3	.1052
	Task Content	5	.0069
2-way Effect	Complexity*Difficulty	4	.0133
	Content*Complexity	6	.0029
	Complexity*Specificity	7	.0013
	Difficulty*Specificity	8	.0012
	Content*Difficulty	9	.0005
	Content*Specificity	14	.0001

These ANOVA results suggest that task complexity and task difficulty explain the majority of the variance in the WD statistic with some effect of task specificity as well. It is interesting to note that modifying tasks so that they are more or less difficult (i.e. modifying the $1 - s_{tk}^{(1)}$ and $g_{tk}^{(1)}$parameters) has an average effect on the WD statistic (median $\eta^2 = .28$) that is essentially identical in magnitude to making changes to the complexity of the game (i.e. changing the number of SKIVE elements required for each task throughout the game; median $\eta^2 = .28$ also).

3.1.3 Follow-up Graphical Analyses. Figure 8 shows the distribution of the WD statistic for individual learners for games designed according to design matrix 1; the distribution is computed across all learner types but is broken down by task difficulty and specificity conditions. Recall that the numbers 1-7 are used to indicate the seven task difficulty conditions as shown in Table A-I and that A, B, C are used to indicate the three task specificity conditions as shown in Table A-II.

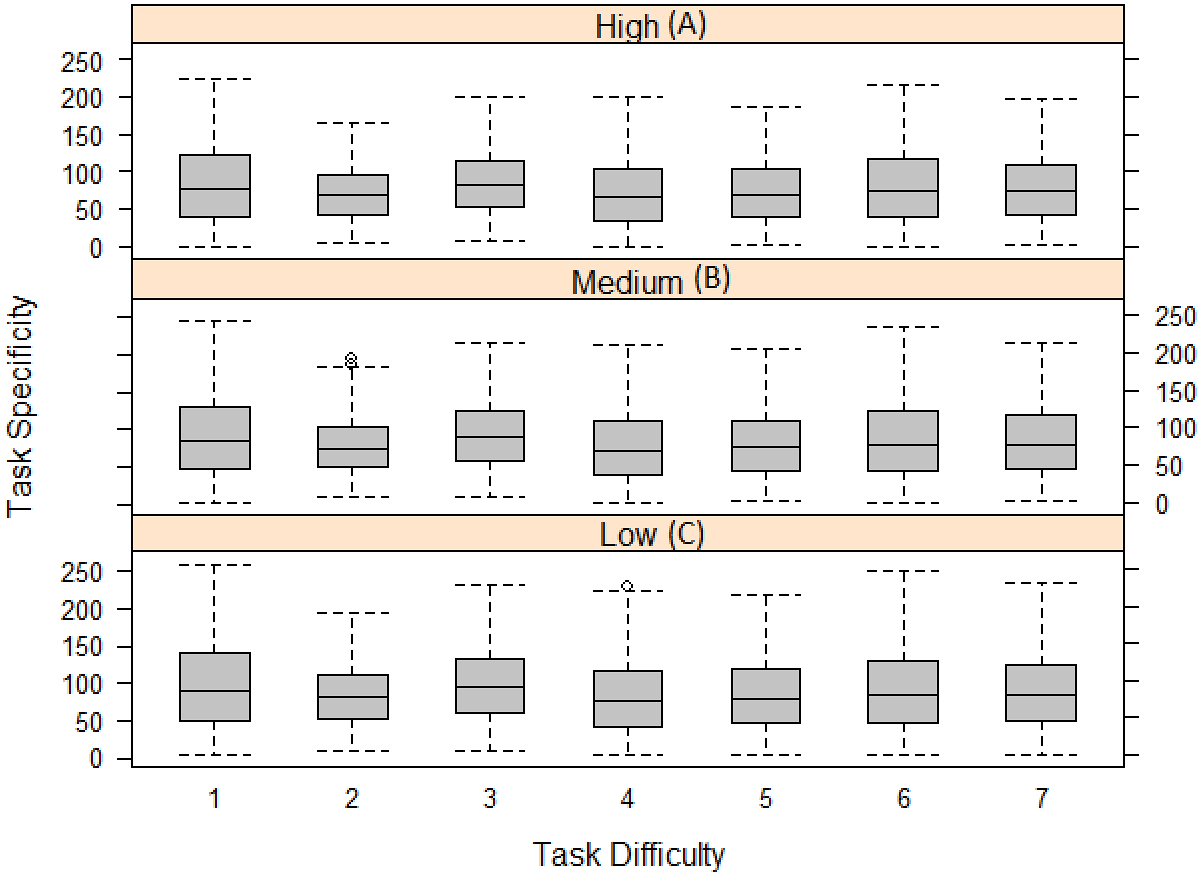


Fig. 8. Distribution of WD statistic for individual learners across learner types for games designed with design matrix 1.

Clearly there is quite a bit of variation in WD values across task difficulty and specificity conditions. Furthermore, in accordance with Table IV, there are fewer differences in variation across task specificity conditions than across task difficulty conditions for games with this design matrix. Notably, there is slightly less variation in conditions associated with "easy" tasks (conditions A2, B2, and C2) and slightly more variation in conditions associated with "difficult" tasks (conditions A1, B1, and C1) with other conditions falling somewhere in between.

The smallest distributional range is observed when the tasks are "easy" as well as "highly specific" (condition A2). In these games all learner types are likely to provide evidence of mastery of targeted

SKIVE elements, independent of whether they have actually mastered them or not. Moreover, learners who have mastered non-targeted SKIVE elements are unlikely to use them since the task parameters prevent such unnecessary use. Consequently, both novice and more expert learners are more likely to generate identical data strings. In contrast, the widest range of WD values is observed for "well-designed" tasks (conditions A1, B1, and C1); remaining conditions represent middle grounds between these.

There is also a slight increase in the positive skew of the distributions across task specificity conditions (i.e. from condition A to B to C, independent of task difficulty settings). This reflects the fact that less specific tasks do not encourage learners to suppress the use of SKIVE elements thereby encouraging the use of unnecessary elements, which, in turn, increases values of the WD statistic. In other words, a combination of easy and less specific tasks, which generally leads to larger numbers of '1's in cumulative adjacency matrices, will lead to more variance in the WD statistic across learners with different developmental trajectories.

The distributions in Figure 8 were computed across all learners with all different developmental trajectories; hence, we subsequently broke these distributions down by learner type. Figure 9 on the next page shows the distributions of the WD statistic for the six learner types with prototypically distinct curvilinear and linear trajectories that we highlighted earlier in Figure 7 for games designed with design matrix 1. As in Figure 8, the distributions are broken down by task difficulty and task specificity conditions.

Overall, it is noteworthy that the boxplots for learner types with distinct trajectories are well separated across task difficulty and specificity conditions. The overlap of the distributions mimics the similarity of the WD traces shown in Figure 7. That is, expert and "faster" learners have more similar distributions and the distributions for the learner types with the three distinct curvilinear trajectories are clearly separated in each game condition. The differential ranges of the distributions that we observed in Figure 8 above are reflected in Figure 9 as well. In conditions when the ranges of the WD distributions are wider – in particular in conditions with a '1' as noted above – the distributions are separated best (i.e. learner differences are most pronounced).

For real data the learner trajectories are of course not known. However, the results shown here suggest that learners with distinct learning trajectories will have ranges of values of the WD statistic that are reasonably distinct. Therefore, a descriptive inspection of the distribution of the WD values for all learners may provide some quick indication about learner differences. For example, when computed during gameplay, the WD statistic could be used as a quick screening tool to suggest pairings of learners who have likely different profiles for subsequent activities of the game (e.g. by pairing learners with lower and higher values of the WD statistic). More subtle nuances in developmental trends will of course not be captured by the statistic.

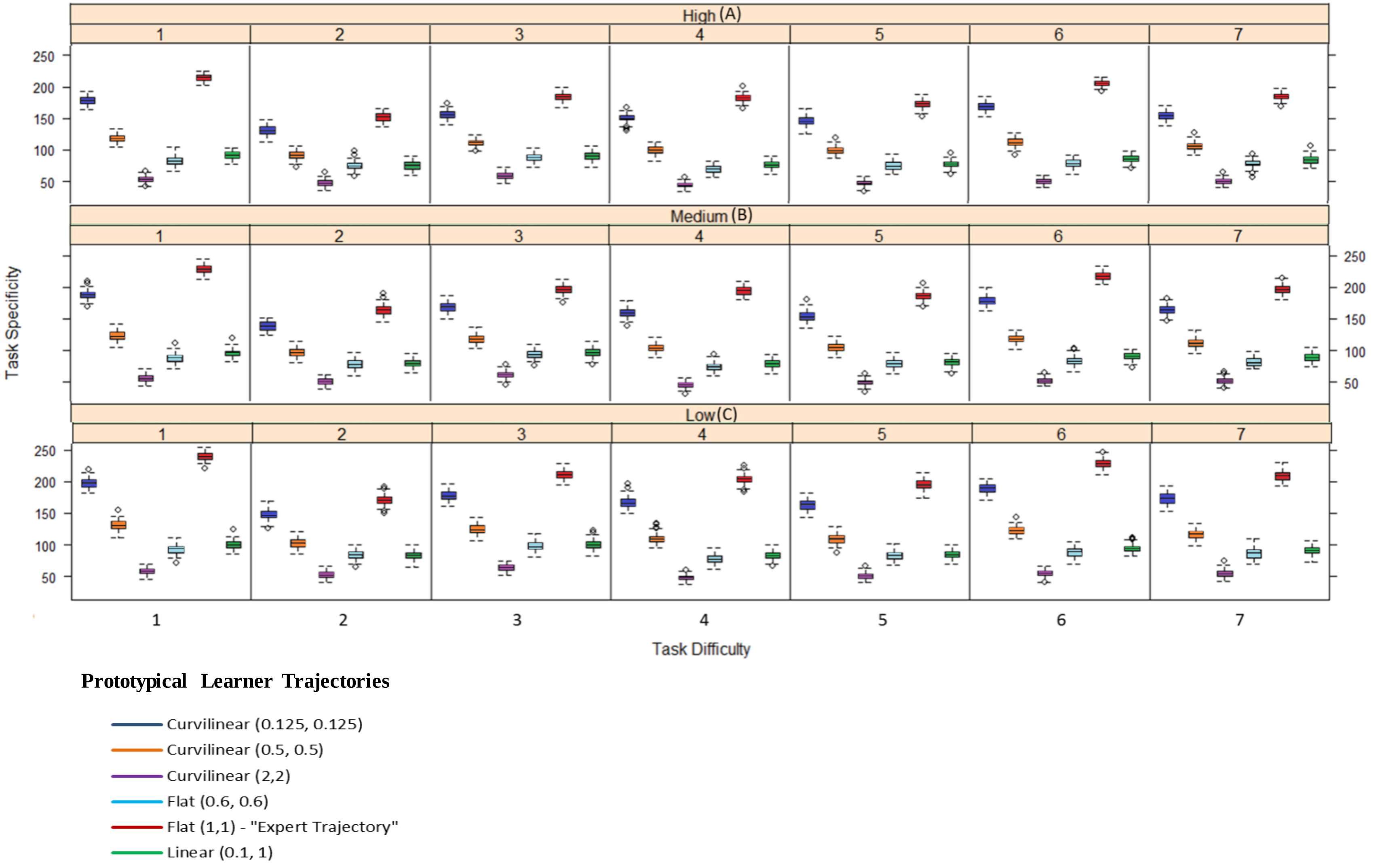


Fig. 9. Distribution of WD for six learner types for games designed with design matrix 1.

3.2 WD Statistic for Pairs of Learners (Variant 2)

3.2.1 General Patterns. Figure 10 shows the WD values for individual learners calculated for the same six learner types with prototypical trajectories shown earlier in Figure 7 who play games designed with design matrix 1. Each solid line again depicts the median WD calculated at each evidentiary segment / task across all 100 replications while the dotted lines now show the 95% empirical pseudo-confidence bands for these learners.

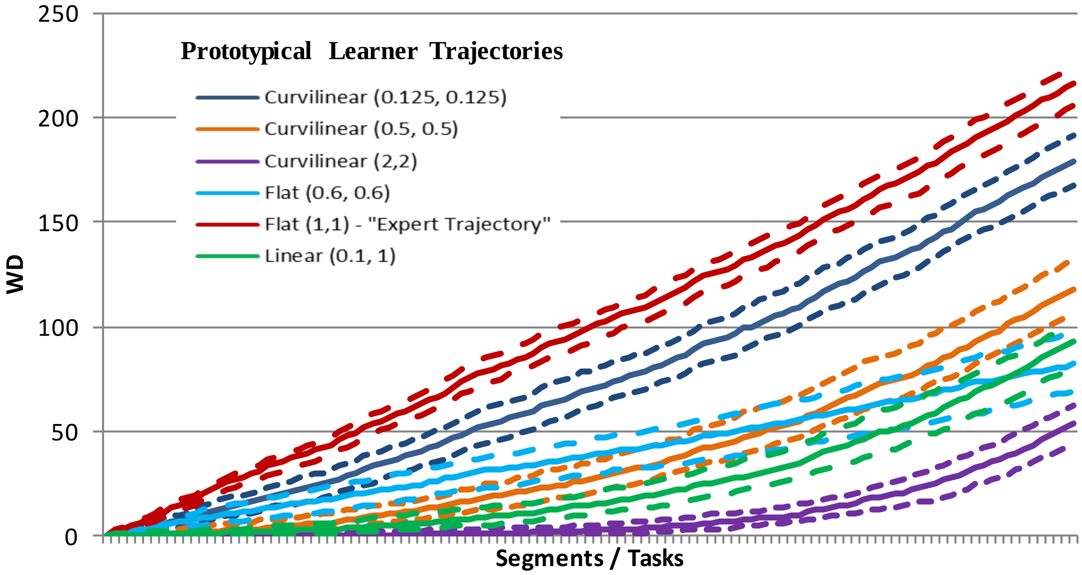


Fig. 10. Empirical 95% pseudo-confidence bands for WD statistic of six learner types for games designed with design matrix 1.

The width of the pseudo-confidence bands increases slightly across segments / tasks, but five out of the six learners – the ones who make actual learning gains throughout the game (i.e. the ones who do not have a flat trajectory) – remain rather well separated after a few evidentiary segments / tasks. As noted in the methods section, the pseudo-confidence bands were used to determine whether the WD distributions for pairs of learners overlapped at each segment / task ('1' if they did, '0' if they did not). The percentage of segments / tasks for which there was overlap was then recorded for each learner pair.

3.2.2 Classification of Learners and ANOVA Results. As with the patterns for the WD statistic for individual learners described in the previous subsection, we submitted the percentage-overlap values of the WD statistic to a factorial ANOVA. With 21 learning trajectories there were a total of $\binom{21}{2} = 210$ different pairs of learning trajectories, however. Consequently, the factor 'type of learning trajectory pair' would have had 210 levels making meaningful comparisons of specific cell means difficult, even with tools such as orthogonal post-hoc contrasts.

Hence, we decided to classify pairs of learner trajectories according to whether they produced "similar" mastery probabilities throughout the game. Specifically, each pair of learning trajectories was coded as follows to create three *learning trajectory similarity groups*; see Tables A-III and A-IV for details on the resulting classifications across trajectories:

(1) 'very similar' if at least 75% of the evidentiary segments / tasks for the two trajectories had mastery probabilities that were within .20 of one another (N_1 = 39 trajectory pairs),

(2) 'moderately different' if between 25% and 75% of the evidentiary segments / tasks for the two trajectories had mastery probabilities that were within .20 of one another (N_2 = 73 trajectory pairs), and

(3) 'very different' if at most 25% of the evidentiary segments / tasks for the two trajectories had mastery probabilities that were within .20 of one another (N_3 = 98 trajectory pairs).

This recoding scheme reduced the original set of 210 trajectory levels to a mere three trajectory group levels, which allowed us to make more meaningful interpretations within the context of a factorial ANOVA.

We note that the classification choices were relatively arbitrary. We thus conducted a series of robustness analyses to investigate to what extent the classification of each pair of learner types and associated interpretational narratives would change if the coding scheme was altered. We calculated the relative similarity of trajectories using probabilities ranging between .05 and .35, rather than just .20. When we used .35, the similarity classifications began to obscure important practical distinctions between learners that gain expertise over the course of game play and an expert player. Using probability values of .05 and .10 resulted in essentially collapsing the three categories into two because so few trajectories were classified as being "very similar." For intermediate values, most notably the classification of learners with flat trajectories changed, but these changes in classification did not alter the key findings of the analysis reported below. For the purposes of the following description, therefore, the recoding scheme as described above was used.

We first conducted a five-way ANOVA with the percentage-overlap WD values as the outcome variable and learning trajectory group (3 levels), task difficulty (7 conditions), task specificity (3 levels), task content (2 levels), and task complexity (2 conditions) as design factors for a total of 252 cells or design conditions. Recall that a set of 100 replications resulted in a single percentage overlap value for the WD statistic so that the number of observations in each cell in the ANOVA design was the number of learner pairs in a particular trajectory group.

The results showed that differences in learner trajectories dominated all other sources of variation (η^2 = 57.90 %) with a decreasing trend in mean percentage-overlap values across the three groups; all of the remaining effect size values were essentially zero. This is clearly expected given that this variant / aggregation of the WD statistic is designed to pick up differences between pairs of learners with distinct trajectories.

However, it was not immediately clear how strong the influence of the game design factors on the variation of the percentage-overlap values was when conditioning on the trajectory similarity. We thus conducted three separate four-way ANOVAs, one for each trajectory group; results are shown in Table V with median η^2 values for main effect and two-way interaction effects shown similar to Table IV. Contrary to the results for the WD statistic for individual learners, only a negligible amount of variation in the percentage-overlap values was accounted for by the game design factors. As we discuss in the next subsection though, this pattern is interpretable.

Table V. Median η^2 for Factorial ANOVA for Version 2 of WD Statistic

Effect	Factor	Very Similar (N_1 = 45)	Moderately Different (N_2 = 75)	Very Different (N_3 = 90)
Main Effect	Task Specificity	.0008	.0043	.0008
	Task Complexity	.0009	.0039	.0034
	Task Difficulty	.0224	.0197	.0179
	Task Content	.0004	.0004	.0005
2-way Effect	Content*Complexity	.0001	.0000	.0000
	Content*Difficulty	.0004	.0005	.0002
	Content*Specificity	.0001	.0001	.0001
	Complexity*Difficulty	.0071	.0026	.0016
	Complexity*Specificity	.0000	.0019	.0001
	Difficulty*Specificity	.0008	.0008	.0004

3.2.3 Follow-up Graphical Analyses. An inspection of the distributions of percentage-overlap values provided some insight into the non-significance of the game design factors observed in Table V. Figure 11 on the next page shows the distributions of the percentage-overlap values for games designed with design matrix 1; as in Figures 8 and 9 the distributions are broken down by task difficulty and task specificity conditions.

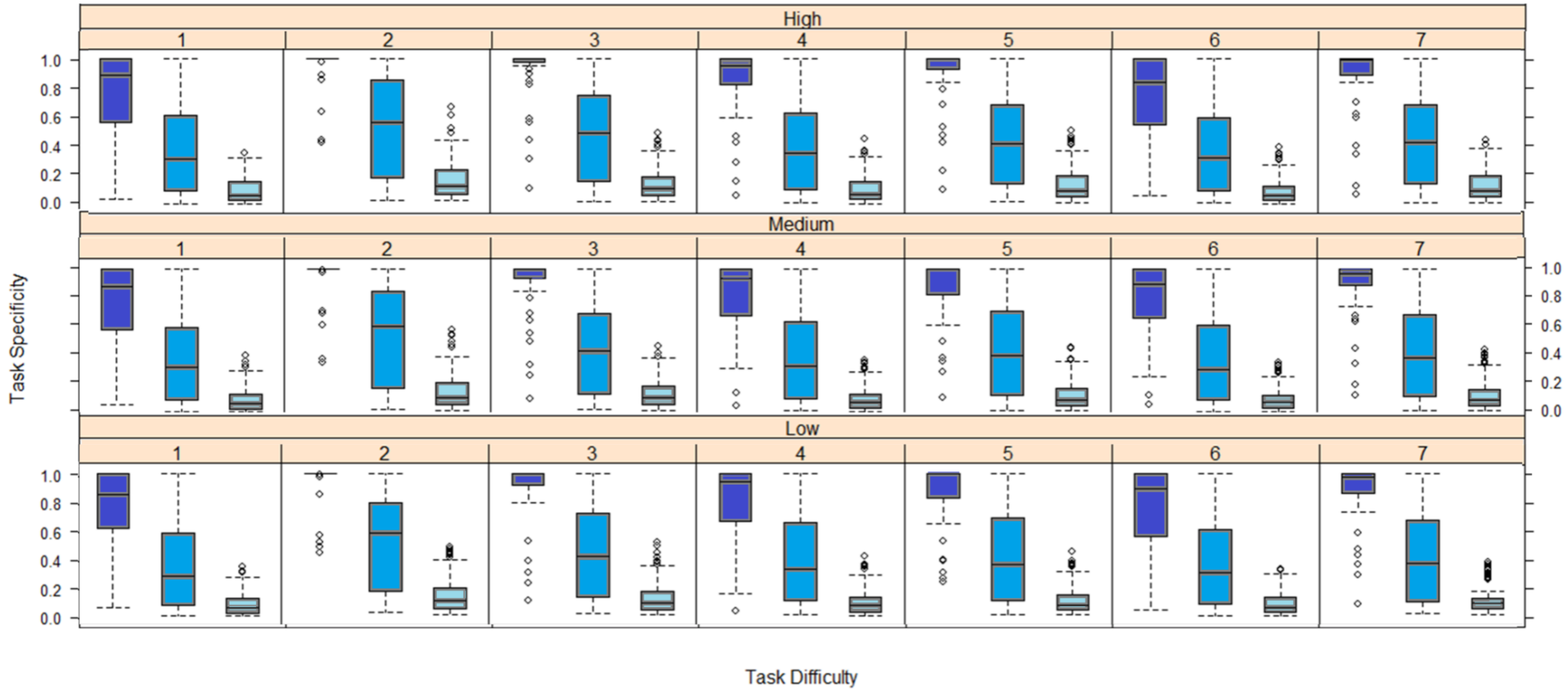


Fig. 11. Distribution of percentage-overlap of WD statistic for learner pairs across trajectory groups and game conditions for a game designed with design matrix 1. Trajectory groups within each cell are ‘very similar’, ‘moderately different’, and ‘very different’ from left to right.

Clearly, the relative locations and spreads of the distributions (i.e. the degree to which they overlap) across the three trajectory groups is as desired in that learner pairs with trajectories classified as ‘very similar’ show higher percentage-overlap values than learners who are classified as ‘moderately different’, who, in turn, show higher values than learners who are classified as ‘very different’. The relative maxima of these distributions are highest for task difficulty conditions 2, 3, 5, and 7, and lowest for conditions 1, 4, and 6. This pattern is consistent across games comprised of tasks that are “highly constrained” (conditions A), “moderately constrained” (conditions B), or “relatively unconstrained” (conditions C). Thus, the same game design conditions that lead to the most reliable differentiations between learner types using the WD statistic also lead to the most reliable differentiations using the percentage-overlap variant of the WD statistic.

4. Discussion

In this section we briefly (a) review the key findings of the study, (b) discuss the generalizability of the results, and (c) reflect on the utility of the ECD framework for guiding simulation work for digital learning environments.

4.2 Critical Review of Findings

This simulation study showed that there is some utility in using the WD statistic for understanding the performance of learners who play epistemic games with different design characteristics. Specifically, task difficulty, task specificity, and task complexity were important factors that affected the performance of individual learners. Moreover, learners who had different developmental trajectories for the SKIVE variables (i.e. the student model variables) could be identified as distinct when using either variant / aggregation of the WD statistic. When the percentage-overlap values were used, distinct learner pairs could be identified relatively consistently independent of which kind of game they played.

Clearly, though, the WD statistic was not equally sensitive across different game conditions. Because it is generally driven by the overall number of ‘1’s in the cumulative adjacency matrix, the WD statistic becomes more sensitive to trajectory differences when students are able to use SKIVE elements that are not essential for the solution strategy. Similarly, differences between learners who play the same game will always be larger for learners who have more distinct trajectories, especially when these learners play games with tasks that are easier and less specific and require multiple SKIVE elements.

Apart from providing some empirical quantification of suspected trends, the design of the simulation study also had some professional development implications for the game development team. It certainly illustrated to them how a distinct modeling of task and learner characteristics resulted in patterns of the WD statistic that were meaningfully interpretable, even if they already knew that the WD statistic was a relatively “blunt” measurement instrument for capturing learners’ emergent expertise. This had implications for how they would think about the game design moving forward in that they could think more critically about fine-tuning tasks to target specific SKIVE elements, for example.

The results also reiterated that a simple marginal summary statistic cannot disentangle the separate influences of task and learner characteristics on game performance for which more sensitive tools are needed. Even though the design team had observed this with real data and was already working on additional tools by the time this study was completed, the simulation study provided a more comprehensive look at the systematic interaction between task and learner characteristics and the direction and magnitudes of resulting effects.

4.2 Generalizability of Results

Our operationalization of game design and learning trajectory conditions as well as our decisions about which kinds of conditions to investigate were necessarily incomplete and somewhat of an over-simplification viz-a-viz real-life game play. For example, the macro-level segmentation of this game into three blocks and the specification of these elements was not without debate among the different project teams involved in this study. On the one hand, this three-stage progression mapped onto the structure of the version of *Land Science* that was in use at the time the simulation was designed, with a series of tasks that included familiarization with the game environment first, a series of tasks completed within homogeneous stakeholder groups next, and a series of tasks completed within heterogeneous stakeholder groups at the end.

On the other hand, different emergent versions of *Land Science* that were created while the study was conducted resulted in slightly different activity structures and specifications of the targeted SKIVE elements. Thus, a different choice for representing game progression may be more appropriate once the design specifications for this and other epistemic games under development are fully formalized and can be reflected in future simulation studies. Future work will need to map the game design specifications from a more stable version of *Land Science* onto a simulation design framework. If attractive, the simulation approach that we have chosen here has certainly enough flexibility to accommodate a wide variety of game design structures.

Similarly, the learning trajectories could be varied across the different SKIVE variables. Furthermore, while the specification of variable-specific trajectories implied an impact on co-occurrences of these variables, we did not model these associations directly. During the design of the simulation study, the game development team noted that a direct modeling of co-occurrences would be a more natural fit for their theory of task engagement via the epistemic frame theory, which could be done in future work.

Yet, even though we were able to conceive of some data-generation mechanisms for co-occurrence structures (e.g., via a tetrachoric correlational model for sets of game segments / tasks with constraints imposed on the variance-covariance matrix), there was no single approach that would have been an immediately clear and superior candidate. Nevertheless, we would argue that any data-generation mechanism for a simulation study for epistemic games should incorporate separate influences of learner and task features on observable performance; our results clearly showed that this generated results that were interpretable and generated insight for all team members.

Data that are being generated with the mechanism implemented in this study – or any other suitable mechanism for that matter – could also be used as input into alternative statistical methods such as multidimensional scaling techniques or methods that use projection methods based on singular value decomposition, for example. Indeed, one promising current variant of ENA uses a principal components analysis for ordinal data to project learner trajectories into a two-dimensional space [d'Angelo, Shaffer, and Hatfield 2011].

Finally, we want to note that we learned an additional important lesson in this project, namely that the complexity of the data structures resulting from our simulation study made the choices for how to aggregate and represent results non-trivial. While it was easy to conceive of a variety of game design and learner development conditions, the number of interactions between these factors increased quickly making effective summaries challenging to construct. In general, we would expect that the data-management and presentation issues that we encountered will not become simpler even when the simulation design is altered.

4.3 Utility of the ECD Framework for Simulation Study Design

The use of the ECD framework to navigate the design process for this simulation study was critical to recognizing both the strengths of design choices and their inherent limitations. For example, understanding that SKIVE elements are student model variables in the student model, that the way tasks measure them as specified in design matrices and task parameters is part of the specification of the task model, and that the binary coding of keywords for SKIVE elements and the associated computation of the WD statistics constitute evidence identification and accumulation processes in the evidence model was insightful. Specifically, the use of the ECD framework and associated terminology helped to operationalize, and disentangle, the constituent drivers of the relatively complex interactions of learners with each other and with the game environment. It also helped to see where the simulation design and the game design were (mis)aligned.

In the end, simulation studies for epistemic games are collaborative endeavors and final designs reflect a series of compromises that typically do not satisfy every member of every team. Teams need to make design decisions that are both statistically useful – in that they induce variation in the statistics under investigation – and practically useful – in that they do this over design conditions that have reasonable degrees of face validity; this is a very difficult balance to achieve. This tension is likely faced by many researchers who conduct simulation work within the context of an evolving digital learning environment.

As we have illustrated in this paper, the ECD framework can offer a much-needed conceptual framework for facilitating communication and joint decision-making between game developers and statisticians because it provides a common language and associated methodological toolbox. Our paper was not designed to provide definitive statistical answers per se but, rather, to provide a leverage point for understanding the complexity of this line of work. We sincerely hope that it was useful to read as a starting point for constructive discussions and reflections in interdisciplinary teams concerned with games-based assessment. We look very forward to a continued dialogue with colleagues in the field that builds upon our initial experiences in this paper and related experiences in similar projects.

ACKNOWLEDGMENTS

This work was made possible, in part, by 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. We want to thank specifically David W. Shaffer and his team for providing critical reviews on the strengths and limitations of this simulation study. We also want to thank three anonymous reviewers and the editors for their constructive feedback on earlier versions of this manuscript, which is greatly improved because of them.

REFERENCES

BAGLEY, E., AND 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.

BAGLEY, E., AND SHAFFER, D. W. 2010. *Stop talking and type: Mentoring in a virtual and face-to-face environment*. Manuscript under review.

BORG, I., AND GROENED, P. J. F. 2010. *Modern multidimensional scaling: Theory and applications*. Springer, New York, NY.

COOK, A. R., AND TEO, S. W. L. 2011. The communicability of graphical alternatives to tabular displays of statistical simulation studies. *PLoS ONE* 6, 1-7.

D'ANGELO, C., SHAFFER, D. W., AND HATFIELD, D. 2011. *Design, implementation, and data analysis for an epistemic game using a novel variant of epistemic network analysis.* Manuscript submitted for publication.

DE AYALA, R. J. 2009. *Principles and Practice of Item Response Theory*. Guilford Press, New York, NY.

MISLEVY, R. J., STEINBERG, L. S., ALMOND, R. G, AND LUKAS, J. F. 2006. Concepts, terminology and basic models of evidence-centered design. D. M. WILLIAMSON, R. J. MISLEVY, AND I. I. BEJAR. Eds. *Automated Scoring of Complex Tasks in Computer-Based Testing*, Lawrence Erlbaum Associates, Mahwah, NJ, 15–47.

NASH, P., AND SHAFFER, D. W. 2011. Mentor modelling: The internalization of modelled professional thinking in an epistemic game. *Journal of Computer Assisted Learning 27*, 173-189.

R DEVELOPMENT CORE TEAM. 2008. *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. Available online at http://www.R-project.org.

RECKASE, M. 2009. *Multidimensional Item Response Theory*. Springer, New York, NY.

RUPP, A. A., TEMPLIN, J., AND HENSON, R. J. (2010). *Diagnostic Measurement: Theory, Methods, and Applications.* Guilford Press, New York, NY.

RUPP, A. A., GUSHTA, M., MISLEVY, R. J., AND 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

SHAFFER, D. W. 2006a. *How Computer Games Help Children Learn.* Palgrave / Macmillan, New York, NY.

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., FRANKI, K., RUPP, A. A., AND MISLEVY, R. J. 2009. Epistemic network analysis: A prototype for 21st century assessment of learning. *The International Journal of Learning and Media 1(2),* 33-53.

TATSUOKA, K. K. 2009. *Cognitive Assessment: An Introduction to the Rule-space Method.* Routledge, Florence, KY.

APPENDIX

Table A-I. Task Parameters Specification for *Activation* of SKIVE Elements for any given Task in a Game

Difficulty	S		K		I		V		E		Description
Condition	$s_{tk}^{(1)}$	$g_{tk}^{(1)}$	$s_{tk}^{(1)}$	$g_{tk}^{(1)}$	$s_{tk}^{(1)}$	$g_{tk}^{(1)}$	$s_{tk}^{(1)}$	$g_{tk}^{(1)}$	$s_{tk}^{(1)}$	$g_{tk}^{(1)}$	
					Same Parameter Pattern for all SKIVE Variables						
1	Low	Low	Low	Low	Low	Low	Low	Low	Low	Low	All tasks are “globally well-designed” (i.e. they facilitate the successful application of mastered SKIVE elements and do not allow for the random application of non-mastered SKIVE elements)
2	High	High	High	High	High	High	High	High	High	High	All tasks are “globally poorly designed” (i.e. they allow for the unsuccessful application of mastered SKIVE elements and the random application of non-mastered SKIVE elements)
3	Low	High	Low	High	Low	High	Low	High	Low	High	All tasks are “globally moderately well designed” (i.e. they facilitate the successful application of mastered SKIVE elements but also allow for the random application of non-mastered SKIVE elements)
4	High	Low	High	Low	High	Low	High	Low	High	Low	All tasks are “globally moderately well designed” (i.e. they allow for the unsuccessful application of mastered SKIVE elements but do not allow for the random application of non-mastered SKIVE elements)

(continued)

Difficulty	S		K		I		V		E		Description
Condition	$s_{tk}^{(1)}$	$g_{tk}^{(1)}$	$s_{tk}^{(1)}$	$g_{tk}^{(1)}$	$s_{tk}^{(1)}$	$g_{tk}^{(1)}$	$s_{tk}^{(1)}$	$g_{tk}^{(1)}$	$s_{tk}^{(1)}$	$g_{tk}^{(1)}$	
Differential Parameter Pattern across Different SKIVE Variables											
5	Low	High	Low	High	High	Low	High	Low	High	Low	Tasks are "differentially well designed" (i.e. for "basic" SKIVE elements they facilitate the successful application of mastered elements but not the random application of non-mastered elements; for more "complex" SKIVE elements they allow for the unsuccessful application of mastered elements but not the random application of non-mastered elements)
6	Low	Low	Low	Low	High	Low	High	Low	High	Low	These tasks are "differentially well designed" (i.e. for "basic" SKIVE elements they facilitate the successful application of mastered elements and do not allow for the random application of non-mastered elements; for more "complex" SKIVE elements they allow for the unsuccessful application of mastered elements but not the random application of non-mastered elements)
7	Low	High	Low	High	Low	Low	Low	Low	Low	Low	These tasks are "differentially well designed" (i.e. for "basic" SKIVE elements they facilitate the successful application of mastered elements but also allow for the random application of non-mastered elements; for more "complex" SKIVE elements they also facilitate the application of mastered elements but do not allow for the random application of non-mastered elements)

Note. Task parameters for each SKIVE element (i.e. student model variable) across segments / tasks are drawn from four-parameter Beta distributions with the following interval boundaries and means. *High:* (.20, .30), mean = .25; *Medium:* (.10, .20), mean = .15; *Low:* (.00, .10), mean = .05. The crossing of difficulty conditions 1-7 with specificity conditions A-C results in 21 task parameter conditions (A1, A2,…, C6, C7).
.

Table A-II. Task Parameters Specification for *Suppression* of SKIVE Elements for any given Task in a Game

Specificity	S		K		I		V		E		Description
Condition	$s_{tk}^{(0)}$	$g_{tk}^{(0)}$	$s_{tk}^{(0)}$	$g_{tk}^{(0)}$	$s_{tk}^{(0)}$	$g_{tk}^{(0)}$	$s_{tk}^{(0)}$	$g_{tk}^{(0)}$	$s_{tk}^{(0)}$	$g_{tk}^{(0)}$	
A	Low	1	Low	1	Low	1	Low	1	Low	1	All tasks are overall "highly specific" (i.e. they do not provide many opportunities to demonstrate mastery of non-targeted SKIVE elements / they strongly facilitate the suppression of non-targeted SKIVE elements)
B	Medium	1	Medium	1	Medium	1	Medium	1	Medium	1	All tasks are overall "moderately specific" (i.e. they provide some opportunities to demonstrate mastery of non-targeted SKIVE elements / they moderately facilitate the suppression of non-targeted SKIVE elements)
C	High	1	High	1	High	1	High	1	High	1	All tasks are overall "not very specific" (i.e. they provide many opportunities to demonstrate mastery of non-targeted SKIVE elements / they weakly facilitate the suppression of non-targeted SKIVE elements)

Note. Task parameters for each SKIVE element (i.e. student model variable) across segments / tasks are drawn from four-parameter Beta distributions with the following interval boundaries and means. *High:* (.20, .30), mean = .25; *Medium:* (.10, .20), mean = .15; *Low:* (.00, .10), mean = .05. The crossing of difficulty conditions 1-7 with specificity conditions A-C results in 21 task parameter conditions (A1, A2,…, C6, C7).

Table A-III. Number and Percent of Trajectories that Each Trajectory is Similar To

	C_0.125_0.125	C_0.25_0.25	C_0.5_0.5	C_2_2	C_4_4	C_8_8	F_0.2_0.2	F_0.4_0.4	F_0.6_0.6	F_0.8_0.8	F_0_0	F_1_1	L_0.1_0.6	L_0.1_0.8	L_0.1_1	L_0.3_0.6	L_0.3_0.8	L_0.3_1	L_0.5_0.6	L_0.5_0.8	L_0.5_1
Very Different	10	7	9	11	6	9	8	7	8	8	6	12	4	9	13	14	9	10	13	13	8
Moderately Different	4	9	9	5	10	6	9	10	6	8	5	7	7	6	3	5	5	5	5	4	4
Very Similar	6	4	2	4	4	5	3	3	6	4	9	1	9	5	4	1	6	5	2	3	8
Very Different	50%	35%	45%	55%	30%	45%	40%	35%	40%	40%	30%	60%	20%	45%	65%	70%	45%	50%	65%	65%	40%
Moderately Different	20%	45%	45%	25%	50%	30%	45%	50%	30%	40%	25%	35%	35%	30%	15%	25%	25%	25%	25%	20%	20%
Very Similar	30%	20%	10%	20%	20%	25%	15%	15%	30%	20%	45%	5%	45%	25%	20%	5%	30%	25%	10%	15%	40%

Note. *C* = curvilinear, *F* = flat trajectory, *L* = linear trajectory. For curvilinear trajectories the numbers reflect exponent values for the polynomial function that creates the trajectory while for flat and linear trajectories they reflect the mastery probabilities at the first and last game segment.

Table A-IV. Classification of Trajectory Pairs according to Similarity

	C_0.125_0.125	C_0.25_0.25	C_0.5_0.5	C_2_2	C_4_4	C_8_8	F_0.2_0.2	F_0.4_0.4	F_0.6_0.6	F_0.8_0.8	F_0_0	F_1_1	L_0.1_0.6	L_0.1_0.8	L_0.1_1	L_0.3_0.6	L_0.3_0.8	L_0.3_1	L_0.5_0.6	L_0.5_0.8	L_0.5_1
C_0.125_0.125																					
C_0.25_0.25	1																				
C_0.5_0.5	2	2																			
C_2_2	3	3	3																		
C_4_4	3	3	3	2																	
C_8_8	3	3	3	2	1																
F_0.2_0.2	3	3	3	2	1	1															
F_0.4_0.4	3	3	2	2	3	3	3														
F_0.6_0.6	3	2	2	2	3	3	3	1													
F_0.8_0.8	1	1	2	3	3	3	3	3	3												
F_0_0	3	3	3	2	2	1	3	3	3	3											
F_1_1	1	2	2	3	3	3	3	3	3	1	3										
L_0.1_0.6	3	3	3	1	2	2	2	1	2	3	3	3									
L_0.1_0.8	3	3	3	2	2	3	2	2	2	2	3	3	1								
L_0.1_1	2	2	1	2	3	3	2	2	2	2	3	3	2	1							
L_0.3_0.6	3	3	2	2	3	3	2	1	2	3	3	3	1	1	2						
L_0.3_0.8	3	3	1	2	3	3	3	2	1	2	3	3	2	1	1	1					
L_0.3_1	2	2	1	2	3	3	2	2	2	2	3	3	2	1	1	2	1				
L_0.5_0.6	3	2	2	2	3	3	3	1	1	3	3	3	2	2	2	1	1	2			
L_0.5_0.8	3	2	1	2	3	3	3	2	1	2	3	3	3	2	2	2	1	2	1		
L_0.5_1	2	1	1	3	3	3	3	3	2	1	3	2	3	3	2	3	2	2	2	1	

Pair Type	*N*
Very Different (3)	98
Moderately Different (2)	73
Very Similar (1)	39

Note. *C* = curvilinear, *F* = flat trajectory, *L* = linear trajectory. For curvilinear trajectories the numbers reflect exponent values for the polynomial function that creates the trajectory while for flat and linear trajectories they reflect the mastery probabilities at the first and last game segment.