Design and Discovery in Educational Assessment: Evidence-Centered Design, Psychometrics, and Educational Data Mining

ROBERT J. MISLEVY
Educational Testing Service
JOHN T. BEHRENS AND KRISTEN E. DICERBO
Pearson
and
ROY LEVY
Arizona State University

Evidence-centered design (ECD) is a comprehensive framework for describing the conceptual, computational and inferential elements of educational assessment. It emphasizes the importance of articulating inferences one wants to make and the evidence needed to support those inferences. At first blush, ECD and *educational data mining* (EDM) might seem in conflict: structuring situations to evoke particular kinds of evidence, versus discovering meaningful patterns in available data. However, a dialectic between the two stances increases understanding and improves practice. We first introduce ECD and relate its elements to the broad range of digital inputs relevant to modern assessment. We then discuss the relation between EDM and psychometric activities in educational assessment. We illustrate points with examples from the *Cisco Networking Academy*, a global program in which information technology is taught through a blended program of face-to-face classroom instruction, an online curriculum, and online assessments.

Key Words: Evidence-centered design, educational data mining, psychometrics, games and simulations, Cisco Networking Academy

Authors' addresses: Robert J. Mislevy, Research and Development, Educational Testing Service, MS 12-T, Rosedale Road, Princeton, New Jersey USA 08541, rmislevy@ets.org; John T. Behrens and Kristen E. Cerbo, Center for Digital Data, Analytics & Adaptive Learning, Pearson Education, 400 Center Ridge Dr., Austin, TX 78753, John.Behrens@Pearson.com and Kristen.DiCerbo@Pearson.com; Roy Levy, School of Social and Family Dynamics, Arizona State University, PO Box 873701, Tempe, Arizona USA 85287-3701, Roy.Levy@asu.edu.

1. INTRODUCTION

Data mining is the process of extracting patterns from large data sets, for purposes that include systems enhancement and scientific discovery [Witten and Frank 1999]. *Educational data mining* (EDM) in particular aims to provide insights into instructional practices and student learning, often using data from assessments and learning experiences, both formal and informal [Romero et al. 2011]. Applying exploratory methods to existing data seems to contrast with forward-design process of developing assessments.

This paper explores the productive dialectic that can be developed between EDM and principled assessment design as seen from the perspective of *evidence-centered design* (ECD) [Almond et al. 2002; Mislevy et al. 2003]. We first present an overview of ECD, with an eye toward complex assessments. We then discuss the relationship between psychometric and EDM activities in assessment, and use the ECD perspective to highlight productive connections between EDM and assessment.

Points are illustrated with brief examples from the literature and from our own work with the *Cisco Networking Academy* (CNA) [www.cisco.com/web/learning/netacad/index.html; see also Rupp et al. this issue]. The CNA is a global program in which beginning computer network engineering and ICT literacy is taught through a blended program of face-to-face classroom instruction, an online curriculum, and online assessments. Courses are delivered at high schools, 2- and 3-year community college and technical schools, and 4-year colleges and universities. Since its inception in 1997, the CNA has grown to reach a diverse population of about a million students each year in more than 165 countries [Murnane et al. 2002; Levy and Murnane 2004]. Behrens et al. [2005] discuss the framework that drives the ongoing assessment activity from which our illustrations are drawn.

2. ASSESSMENT, ECD, AND PSYCHOMETRICS

Most familiar applications of educational assessment are framed in what we will call the *standard assessment paradigm*. Data from each student are sparse, typically discrete responses to perhaps 30 to 80 test items. The items are predefined. The target of inference is a student's level of proficiency in a domain framed in trait or behaviorist psychology and defined operationally by the items. Learning during the course of assessment is assumed to be negligible. We can view the standard assessment paradigm as a subspace of assessment viewed more broadly, where any or all of the familiar constraints could be relaxed: continuous performances in interactive environments, for example; richer data that encompass many aspects of activity at any level of detail; interest in multiple aspects of proficiency, evoked in different combinations in different situations; learning may occur, and may indeed be an aim of the experience.

By definition, psychometrics is measuring educational and psychological constructs. Psychometrics in educational testing has focused mainly on da ta produced in the standard assessment paradigm. Much progress in test theory has been made "by treating the study of the relationship between responses to a set of test items and a hypothesized trait (or traits) of an individual as a problem of statistical inference" [Lewis 1986, p. 11]. Probabilistic test theory models allow an analyst to characterize the informational value of data about students in a probabilistic framework, and to use data from different tasks to draw inferences in terms of the same proficiencies. These are powerful inferential tools for practical work in assessment.

The challenge for educational assessment is to jointly harness EDM capabilities to deal with the richer data environment in which we can now carry out assessment, and the inferential strengths of psychometric methods that have evolved for inference with data from the standard assessment paradigm.

The way forward is an assessment framework that encompasses both perspectives, and supports the design and analysis of both familiar assessments and new ones that take advantage of technological advances to move beyond the standard assessment paradigm. Such a framework would embrace concepts and methods from EDM as well as from existing psychometrics. Recent work in assessment provides a suitable foundation. One line of progress is the conception of assessment as argument [Cronbach 1980, 1988; Kane 1992, 2006; Messick 1989, 1994]. Another is the integration of psychometric modeling, assessment design, and cognitive theory [Embretson 1985; Pellegrino et al. 2001; Tatsuoka 1983]. ECD builds on this research to provide a framework for analyzing and carrying out assessment design and implementation. We will use it to bring out productive and natural roles of EDM in the assessment enterprise.

3. EVIDENCE-CENTERED DESIGN

Assessment in the standard assessment paradigm is thought of just in terms of the highly scripted circumstances in which students solve constrained tasks, usually answering verbal questions, and results that simply accumulate independent item scores. The ECD framework neither requires nor implies this limited view of assessment. It is flexible enough to describe a wide range of activities and goals associated with assessment as conceived more broadly, including familiar tests but accommodating the informal assessment activities of instructors interacting with students in the classroom, students working through open-ended simulation tasks [Frezzo et al. 2009; Mislevy 2011; Williamson et al. 2004], multi-student interactions in role playing or simulated situations [Shute 2011], and game-based assessments [Behrens et al. 2007; Mislevy et al. in press].

ECD emphasizes the specification of the logic of assessment, or the *evidentiary argument* [Embretson 1983]. Messick [1994] describes the structure of an *assessment argument* as follows:

> A construct-centered approach would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics. (p. 16)

ECD formalizes this structure with an explicit framework for designing and implementing assessments. The following section sketches the key concepts and representations. At a given point in time, for some practical assessment purpose, one can use ECD to design the components of an operational assessment. Design is practical, and it is provisional as well. Especially with more complex forms of assessment, we expect our understanding of the nature of proficiency to improve as we explore patterns in data from a given form of the assessment [Behrens et al. 2012]. Bringing EDM tools to bear on the data at any given point in time can thus lead to deeper understanding and improvements for assessment design and analysis in the next version of the assessment.

3.1 Assessment Components and ECD

Figure 1 distinguishes five ECD "*layers*" at which different types of thinking and activity occur in the development and operation of assessment systems [Mislevy and Riconscente 2005, 2006].

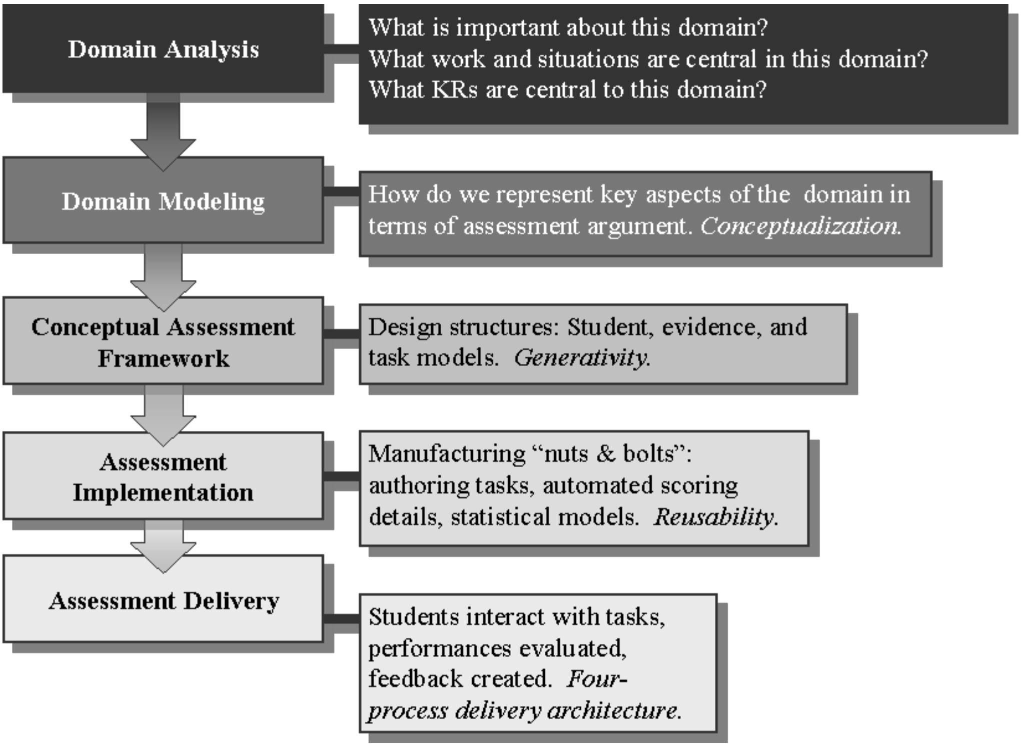
Domain Analysis
What is important about this domain?
What work and situations are central in this domain?
What KRs are central to this domain?
Domain Modeling
How do we represent key aspects of the domain in terms of assessment argument. *Conceptualization.*
Conceptual Assessment Framework
Design structures: Student, evidence, and task models. *Generativity.*
Assessment Implementation
Manufacturing "nuts & bolts": authoring tasks, automated scoring details, statistical models. *Reusability.*
Assessment Delivery
Students interact with tasks, performances evaluated, feedback created. *Four-process delivery architecture.*

Fig. 1. Layers in the evidence-centered assessment design framework.

Each layer contains components, processes and representations that are appropriate for the kinds of activities that take place in that layer (see Mislevy et al. 2010 on the central role of representations in ECD). *Formal object models* [Rumbaugh et al. 1991] have been implemented in the *Portal* design system [Almond et al. 2003; Steinberg et al. 2005] and the *Principled Assessment Design for Inquiry* (PADI) design system [Riconscente et al. 2005]. Although the figure might suggest a linear design and implementation process, iterative feedback loops are essential to successful designs. In the second part of the paper we will discuss critical roles that EDM can play in iterative design in assessment.

Table I summarizes how the assessment layers play out in classroom instruction, standardized accountability assessment, and interactive diagnostic computer systems. The terminology for the layers and their components that appears in the table will be developed as we describe the layers in turn.

Table I. Summary of Layers of the ECD Framework and Conceptualization of Activity from the ECD Perspective for Three Kinds of Assessment

ECD Layer	**Epistemic Focus**	**Classroom Instruction**	**Standardized Accountability Measure**	**Tutoring System**
		Attend to specific strengths and errors while managing administrative requirements.	Broad inference from wide sample of performance.	Inference regarding specific functional states of students' knowledge and skill and providing experiences to improve them.
Domain Analysis	Understand proficiencies, conditions of use, practices, representations, standards, activities, etc. in the targeted domain.	Teacher's background studies of learning and of the curricular goals.	Common texts associated with curriculum; ongoing scientific activity feeding in to standards and practices.	Cognitive task analysis; protocol analysis; literature related to domain.

(continued)

ECD Layer	Epistemic Focus	Classroom Instruction	Standardized Accountability Measure	Tutoring System
Domain Modeling	What are relationships among proficiencies, task situations, and performance in such situations; What is important for the purpose(s) of the assessment?	Teacher's mental model of curriculum elements & dependencies, and situations for learning about students' proficiencies.	Standards documents and assessment frameworks; e.g., Common Core State Standards, National Science Education Standards.	Specifications of production rules and their combinations. Relationship of performances & products to production rules.
Conceptual Assessment Framework (CAF)	What are the linkages between tasks & evidence about proficiency; i.e., what are schemas for tasks, procedures to capture and evaluation performance?	SM: Aspects of student performances worth tracking. TMs: Informal catalog of kinds of tasks for administrative ease as well as verisimilitude to natural tasks (e.g., writing). EM: Theory of what good performance looks like, to be applied on-the-fly. MM: Add up individual items and grades. Weight differentially if desired.	SM: Core dimensions of proficiency aligned with standards. TM: Schemas for tasks to evince correctness of procedure or knowledge. EM: Evaluation procedures for task types (e.g., right/wrong, partial credit, scoring rubrics). MM: Models that maximize precision of latent variable estimate and efficiency of delivery.	SM: Specification of target proficiencies–production rules or aggregates of them needed to guide students' activity. TM: Specification of features of tasks appropriate for different aspects of the learning progression. EM: Procedures to evaluate features of performances and/or products. MM: Fine-grained model such as Bayes net or cognitive diagnosis model.

(continued)

ECD Layer	Epistemic Focus	Classroom Instruction	Standardized Accountability Measure	Tutoring System
Assessment Implementation	Based on existing knowledge, data, and specifications from above, create the elements needed for the assessment.	Create items for quizzes and tests. Establish grade book. Adjust tests to match changes in curricula.	Author specific tasks; develop scoring keys or rubrics; calibrate IRT model; Assemble forms to optimize Test Information Function. (can cycle with field tests and calibration samples)	Build rules into interactive system, for managing information for inference and instruction choices (e.g., Bayes nets, updating rules based on learning theory) Observe in pilot phases and modify rules and system responses.
Assessment Delivery	Create or co-opt circumstances to obtain relevant evidence.	Observe work on classroom assignments & behavior. Update mental and grading models. Check grades against overall impression or specific activity performance. Iterate between global and diagnostic levels.	Deliver common activities (items), perhaps with limited customization such as computerized adaptive testing for efficiency. Paper and pencil or computer based delivery.	Ongoing interaction of students and computer system, with cycles of presentation, student activity, evaluation, feedback and adaptation of learning situation.
Post Assessment Delivery	Communicate inferences and implications	Report card. Ongoing verbal feedback.	Performance report typically in relation to performance of others	Estimates of proficiency and reporting of particular error patterns and progress.

Notes. Within the CAF: SM = student model; EM = evidence model; MM = measurement model.

3.1.1 Domain Analysis Layer. The first layer of the ECD framework is *domain analysis*. Domain analysis marshals beliefs, representations, and modes of discourse for the target domain. This can include best practices, research findings, practitioners' experiences, expert-novice studies, and historical or sociological framings of a set of knowledge and skills. Understanding the *epistemic frame* of a domain [Shaffer 2006] helps a designer avoid confusing the mastery of isolated tasks with functional mastery of work in a domain. It is not simply "the content" of the domain that matters but how people think with that content, what they do, and the situations in which they do it.

3.1.2 Domain Modeling Layer. The second layer of the ECD framework is *domain modeling*. Assessment developers organize insights about the domain from domain analysis into the form of assessment arguments, in representations that more formally reflect the structure of the Messick quote. They articulate structures and dependencies in knowledge, skills and attributes in the domain, and the relationships of these capabilities to situations and activities. Useful representations include standards formulations, scope and sequences in curriculum, concept maps [DiCerbo 2007], hierarchies of skill dependencies or progressions, Toulmin diagrams, and assessment design patterns.

Specifically, *Toulmin diagrams* for assessment arguments map out the relationships among proficiencies, performances, features of work, and features of task situations [Mislevy 2006]. *Design patterns* sketch out a design space for task authors, with options and examples that draw on research and experience with a certain kind of proficiency [Liu and Haertel 2011]. For example, Mislevy et al. [2009] describe a suite of design patterns that help designers create tasks to assess model-based reasoning. The conceptualizations in domain modeling that ground the design of operational assessments can be continually extended and refined as knowledge is acquired in prototypes, field trials, and analyses of operational data.

3.1.3 The Conceptual Assessment Framework Layer. The CAF lays out more formal specifications for the operational elements of an assessment. Designers combine domain information with information about goals, constraints, and logistics to create a *blueprint* for an assessment, in terms of psychometric models, specifications for evaluating students' work, schemas for tasks, and, in technology-based assessments, specifications of the interactions that will be supported. The CAF thus provides structures that bridge work from the domain analysis and domain modeling phases and the actual objects and processes that will constitute the operational assessment, which are the processes described in assessment delivery layer.

The CAF comprises models (as noted above, software-engineering object models in the sense of Rumbaugh et al. 1991) whose objects and specifications provide the blueprint for tasks, evaluation procedures, and statistical models and delivery and operation of the assessment. The following paragraphs describe the central CAF models depicted in Figure 2.

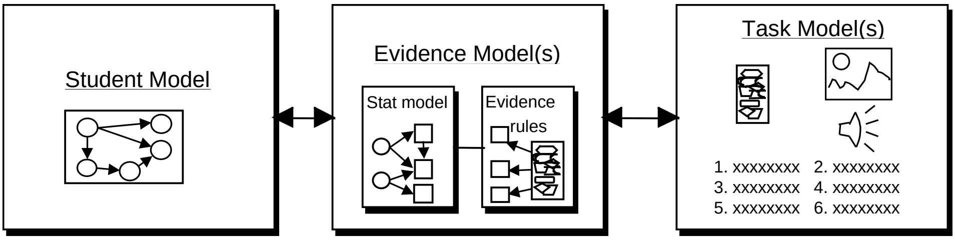

Fig. 2. The central models of the conceptual assessment framework.

A *task model* is a set of assumptions and structures describing task and environment features. Key design elements include the specification of the cognitive artifacts and affordances needed to support the student's activity and the forms in which students' performances will be captured (i.e., work products), such as the sequence of steps in an investigation or the final solution of a design problem. The variables in task models play key roles in assessment arguments, task design, and psychometric models [Mislevy et al. 1999].

The *student model* contains variables for expressing claims about targeted aspects of students' knowledge and skills, at a grainsize and nature that suits the purpose of the assessment. That is, the student model consists of the variables and the structure of those variables in the psychometric model used to synthesize information about aspects of students' proficiencies. It formalizes the aspects of the capabilities identified in the domain model that will be incorporated into the inferential logic of the operational assessment.

The *evidence model* bridges the student model and the task model. It consists of two components: the *evaluation component* (i.e., evidence identification component) provides the rationale and specifications for how to identify and evaluate the salient aspects of work products, which will be expressed as values of observable variables. Data that will be generated in the evaluation component are synthesized across tasks in the *measurement model* component (i.e., evidence accumulation component). The simplest measurement models contain summed scores of salient features of a performance such as the number or percentage correct score. More complicated measurement models such as models from *item response theory* (IRT) [e.g., de Ayala 2009; Hambleton and Swaminathan 1985; Lord 1980; Reckase 2009], *diagnostic classification models* (DCMs) [e.g., Rupp et al. 2010], and *Bayesian networks* (BNs) [e.g., Levy and Mislevy 2004] include formal latent variables. For example, BNs extend the concept maps used in domain analysis to support probabilistic inference [Jensen 1996], such as modeling student skill levels on learning progressions [West et al. 2010]. Although the statistics behind BNs can be complex, the graphical displays that represent the statistics are more accessible to task designers, teachers, and students [DiCerbo 2009].

It is in the CAF, and specifically in the student model and measurement model component of the evidence model, that the previously mentioned two insights of psychometrics in the standard assessment paradigm are implemented - characterizing the weight of evidence in a formal probability model, and enabling for evidence from different tasks to be synthesized in terms of evidence about the same latent variables for student proficiencies. These ideas are so central to the interplay between psychometrics and EDM that we will devote a section to them after the survey of the ECD framework.

As other articles in this special issue demonstrate, simulation and game-based assessments are a developing frontier of assessment [see also Rupp et al. 2010; Shute 2011]. This is especially the case for the evaluation and measurement components of the evidence model. As we noted above, most of the existing practices and the language of measurement evolved for tests consisting of discrete, pre-packaged, tasks with just a few bits of data. Measurement researchers are extending the evidentiary reasoning principles that underlie familiar test theory for this kind of data to the new environment of the "digital ocean" of data [DiCerbo and Behrens 2012; Junker 2011]. It is these rich, complex, and interactive contexts in which EDM will be most valuable.

3.1.4 Assessment Implementation Layer. The fourth layer of the ECD framework is the *assessment implementation* layer. In this layer, assessment practitioners create functioning realizations of the models articulated in the CAF. Field test data are used to check model fit and to estimate parameters of the operational system. The data structures of tasks and parameters are in the forms specified in the CAF models. Some tasks may be omitted from subsequent consideration because of unexpected interactions with student characteristics, misinterpretation, or other functional issues.

Assessment implementation interacts with other ECD layers at this point, in two directions: moving down the layers, toward operation, the data from field tests are used to tune and parameterize tasks and scoring algorithms. Moving back up the layers, unanticipated results and new discoveries can lead to improvements in the CAF models, further back up to new forms for the elements of assessment arguments, or even further back to fundamental advances in understanding of the domain. The logic of these iterations is similar in many respects to the iterative logic of exploratory data analysis [Tukey 1977; Behrens 1997; Behrens et al. in press]. As in EDM and exploratory statistical analysis, there is a stance of skepticism, and “multiple-working hypotheses” are iteratively reduced through detailed data display, model fit analysis and sensitivity analysis.

3.1.5 Assessment Delivery Layer. The fifth layer of the ECD framework is the *assessment delivery* layer. In this layer, students interact with tasks, their performances are evaluated, and feedback and reports are produced. Almond et al. [2002] lay out a *four-process delivery architecture / four-process model* that can be used to describe delivery processes and associated infrastructure components for assessments that range from computer-based testing procedures, paper-and-pencil tests, informal classroom tests, tutoring systems, and one-to-one tutoring interactions. The processes are thus defined in terms of activities and information, and could be carried out by computers or humans or some combination, and the architecture is indifferent to the implementation. Behrens et al. [2008] show how the logic and the structure of this architecture can be extended to games.

Figure 3 shows the principle processes and their interconnections. Next to each process are additional symbols indicating relevant data types available in complex systems. The activity selection process creates an appropriate task or activity, or selects one in light of what is known about the student. The presentation process interacts with the student and captures work products. The evidence identification process is variously called response processing, feature identification, or task-level scoring. This process evaluates work products by methods specified in the evaluation component of the evidence model. This process sends values of observable variables to the evidence accumulation process, or test-level scoring, which uses the measurement models to summarize evidence about the student model variables and produce score reports.

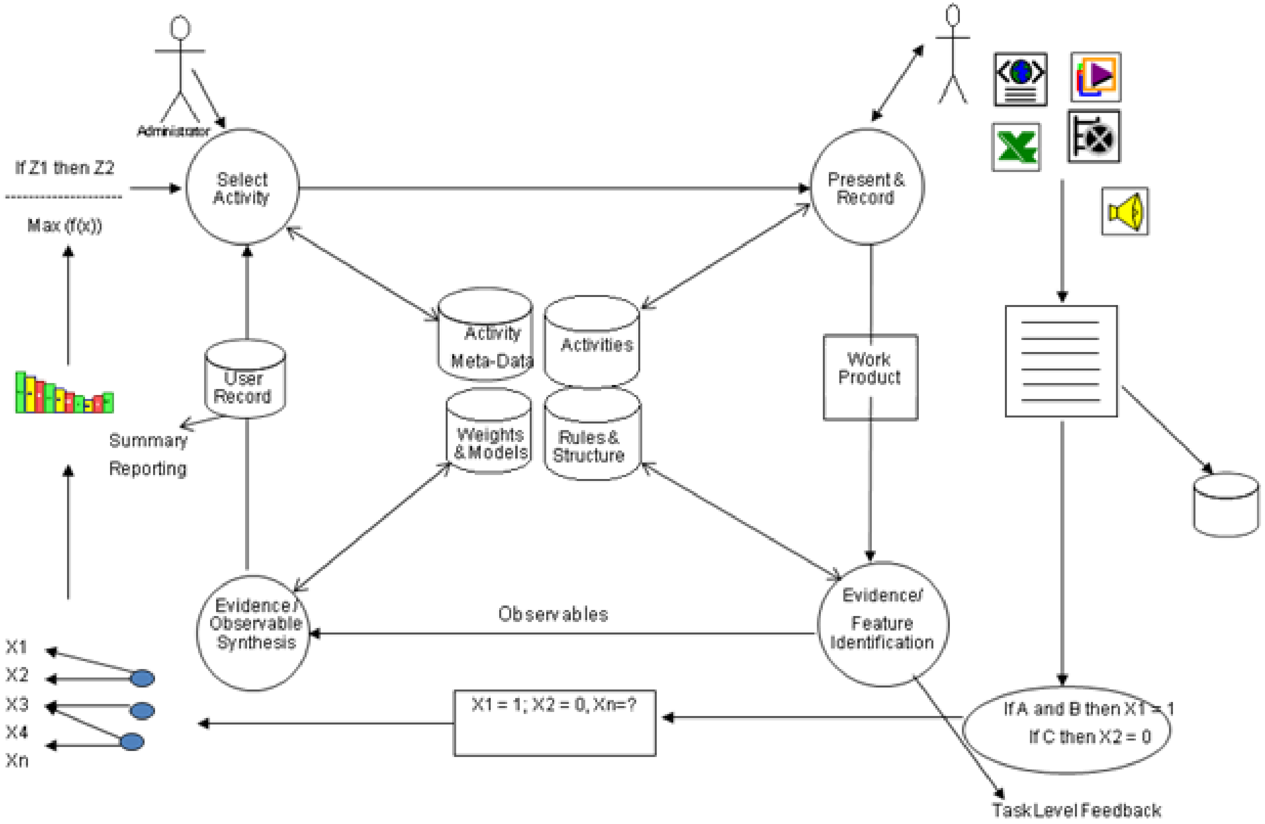

Fig. 3. High-level view of the four-process delivery architecture.

When an assessment is operating, the processes pass messages in a pattern determined by the test's usage - different patterns of scoring, student interaction, and reporting are employed for formative tests, interactive simulations, and batched tests for state-level surveys, for example. The messages are data objects (e.g., parameters, stimulus materials) or are produced by the student or other processes in data structures (e.g., work products, values of observable variables) specified in the CAF [see Almond et al. 2002, for details on the relationships, and Almond et al. 2001, for simple worked-through examples].

Both evidence identification and evidence accumulation are fertile grounds for EDM. For evidence identification, the challenge is finding, combining, and characterizing salient bits of information as features of often-complex work products. This activity does not need to be limited to simple matching but can come from a broad range of symbolic or statistical computations [Williamson et al. 2006]. Automated scoring of spoken responses, for example, can consist of multiple stages, from acoustic analysis, to extraction of features using natural language processing, to statistical combinations of features to produce scores for various aspects of the performance [Bejar 2010]. For evidence accumulation, the challenge is determining useful ways of combining, interpreting, and drawing inferences from these features. Discoveries feed back as improvements to the evidence model in the CAF. Insights into what is important to observe give us better ideas on how to evoke evidence and produce work products, which feed back to the CAF as improved task models.

Distinguishing the processes of a delivery system brings to light conceptually distinct activities in assessment that are obscured in multiple choice testing. Standard practice tightly binds the presentation format (multiple-choice items), work products (mark an option), evidence identification (matching the marked option with the key), and evidence accumulation (count the number of correct responses). The articulated architecture emphasizes that the purpose of the presentation activity is to elicit a work product that could be a simple choice, but could be a complex result such as an essay, activity log, or complex outcome found in the business world (proposal, spreadsheet, or video).

The interaction among the four processes for a fixed-form paper-and-pencil multiple-choice test is a single trip around the cycle if both the content and ordering of tasks is fixed. In a computer-adaptive test that maximizes information about each student individually, the cycle around all four processes occurs for each item: an item is presented and the work product, namely the response, is obtained. Evidence identification evaluates its correctness and passes the result to evidence accumulation. Evidence accumulation updates belief about the student’s proficiency, using for example an IRT model. This information is passed to activity selection, which selects the next item to be most informative, in light of responses up to this point [Wainer et al. 2000; van der Linden and Glas 2010].

A simulation-based task can require many interactions among the processes. Frezzo et al. [2009] describe the interplay among the four processes in the context of the CNA's simulation-based *Packet Tracer Skills-Based Assessment* [see also Rupp et al., this issue]. Activity selection is currently done largely outside the simulation system, although in the articulated architecture this can be changed with minimal changes to the other processes. In the presentation process, the simulation and visualization affordances of the *Packet Tracer* tool allow for presenting tasks that include a broad range of networking devices and protocols. A variable manager allows for the random, or otherwise algorithmic, generation of specific values of features in the environment from lists or numeric ranges. Next, in order to evaluate the work products, *Packet Tracer* provides task authors with a comprehensive list of network states, lets them select which low-level work product features to use in scoring, then allows them to craft scoring rules to apply to these features to create observables.

Note that observable variables need not be answers to discrete, pre-packaged questions, but rather identification of salient features in recurring situations within a continuous flow of performance. For example, using an ECD framework, DiCerbo and Behrens [2012] argue that as daily activity becomes increasingly digital the separation of activity for assessment or non-assessment purposes can be reduced since digital environments are often naturally instrumented to collect work products unobtrusively. When considering digital work products, the assessment designer is faced with the challenge of encoding features of the work product into observations that can be synthesized in the evidence accumulation process. This is leverage point for EDM because it concerns pattern recognition and dimension reduction.

Consider for example, Theodoridis and Koutroumbas's [1999] generic pattern recognition process shown as Figure 4.

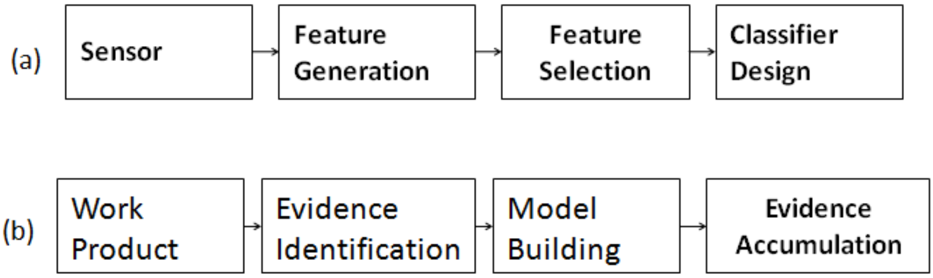

Fig. 4. (a) Pattern recognition process as described in Theodoridis and Koutroumbas [1999]. (b) Corollary to scoring and inference process as described in the ECD literature.

Its steps can be related to the delivery processes of the ECD framework. Output from the sensor in this model is equivalent to the work product produced by the presentation process. Feature generation is concerned with the creation of variables that can be used to describe aspects of the data, corresponding to observable variables in ECD. Feature selection is concerned with determining which observable values are useful for the evidence accumulation process (or perhaps task level or diagnostic feedback as well). Classifier design corresponds to the measurement model / psychometrics for classifying students or measuring proficiency.

ECD provides a flexible and abstracted understanding of assessment data and its relationship to the evidentiary assessment argument, and advances in technology provide new opportunities to link work products to inferences about human states through evidence identification and accumulation. However, in many cases there is weak or little theory that can relate variations in complex work product data (e.g., logs, videos, transcripts of group interactions) and inferences regarding student states, so methods for pattern extraction and other EDM approaches will be needed.

4. PSYCHOMETRICS AND DATA MINING

The psychometric paradigm has been the dominant analysis framework for educational assessment for more than a century, although mainly with the sparse data (at the level of each student) that characterized the standard assessment paradigm. The new field of EDM seeks to improve ways of learning and assessing that are beyond the reach of established analytic practice. This section brings out some essential similarities and differences in the approaches.

4.1 The Ontology and Epistemology of Psychometrics

The view that underlies ECD is that assessment is not simply about producing scores but about obtaining evidence about aspects of students' proficiencies, and characterizing the meaning and the value of that evidence [Mislevy 1994]. Psychometricians use particular kinds of statistical models to quantify these arguments, based on the relevant forms and patterns in data. The two key insights mentioned earlier are (a) characterizing the value of evidence about students' proficiencies in a probabilistic framework and (b) using latent variable models to synthesize evidence from different collections of tasks in a common interpretative frame.

The central insight in characterizing evidence is this: there is a difference between what we observe and what we really want to make inferences about, and the features of the observational situation impact the quality of our inferences. *Classical test theory* (CTT) [Gullikson 1961] uses standard errors of measurement and reliability indices to characterize the accuracy of students' test scores. CTT machinery enables researchers to design tests and compare alternative scoring approaches to improve their work [Cronbach et al. 1972].

The meaning of test scores remains closely bound to the particular items that make up a given test, however. Except in special cases, there is no straightforward way to relate performance on one set of test items to another.

Latent variable psychometric models incorporate an additional insight: we can model the capabilities of people and the features of tasks in ways that enable us to draw inferences about people from different observational situations. Latent variable models posit probability distributions for patterns in observed variables as functions of unobservable (i.e., latent) variables that characterize students' knowledge, skills, strategy repertoires, misconceptions, degree of automaticity, or other cognitively relevant aspects of their capabilities.

The variables in such as model are specified in the student model in the CAF of the ECD framework. They are persistent in the sense that they are posited to influence performance across some domain of tasks where the set of proficiencies they characterize is relevant, and performance in any of the tasks provides evidence about these student model variables. Exactly how performance in each task depends on the student model variables is specified in the measurement models of the CAF. This structure allows a stable frame of interpretation across task situations that may differ markedly on the surface. It becomes possible to assemble psychometric models for different situations according to the features of the situations [Rupp 2002].

As with CTT, a latent-variable modeling framework provides a quantitative basis for operational matters as such planning test configurations, calculating the accuracy and reliability of measurement, figuring out how many tasks or raters we need to be sufficiently sure about the appropriateness of decisions based on test scores, or monitoring the quality of large-scale assessment systems. These models can also be applied to new kinds of testing processes, such as simulation-based tasks and game-based assessments [Mayrath et al. 2012]. Rupp et al. [this issue] employ DCMs and BNs for CNA's simulation-based *Packet Tracer Skills-Based Assessment.*

Although both psychometric models and many EDM models are statistical models, there are distinguishing characteristics of psychometric models and the way they are used. First is the psychological construal of the latent variables in psychometric models. Their interpretation may be cast in terms of behavioral, trait, information-processing, or sociocultural psychology [Mislevy 2006], but in all cases they effect some view of aspects of students' capabilities. The information in patterns of data is synthesized as evidence to characterize students in terms of their standing on these latent variables in the student model.

Second, the nature and grainsize of these student model variables is shaped by the purpose of the assessment. Is the purpose to provide broad feedback about students' general level of proficiency? A psychometric model with few variables, perhaps even an IRT model with a single latent variable for overall proficiency in the domain [Lord 1980] and cast in trait theory, might suffice. Is the purpose of the assessment to provide diagnostic information to guide instruction? A more detailed DCM [Rupp et al. 2010] cast in an information-processing cognitive perspective will be better suited. Is the purpose to characterize knowledge and strategy use in interactive problem solutions? A modular BN approach with models assembled on the fly [Shute et al. 2009] that draws on a situative psychological perspective can be pressed into service. The aim is not simply to discover and model patterns in data; it is to model those patterns that are relevant to specific, practical, educational purposes, in terms that directly inform those purposes.

Third is the explicit mathematical separation of observed score variables from latent student model variables. As noted above, the probability distributions of score variables are modeled as a function of student model variables. More important technically is that the observed variables - which may be characteristics of patterns across lower-level data features - from a given task situation are modeled as conditionally independent of data variables from other task situations. When such a model fits data from some domain of tasks satisfactorily, the underlying patterns in performance that are manifest in different raw data in different task situations can be modeled in terms of the same variables in a student model that can be used with different tasks for different students or at different time points.

Computer-adaptive tests use this idea with IRT, so that students get different items, harder or easier, based on how well they are doing [van der Linden and Glas 2010; Wainer et al. 2000]. Cognitive diagnostic tests use different tasks based on the same cognitive features for teaching and testing mathematics skills [Leighton and Gierl 2007]. Iseli et al. [2010], Shute et al. [2009], and VanLehn [2008] use the idea to build BNs on the fly in game-based and simulation-based assessments to harvest evidence about students' skills from the unique situations, as agents recognize cognitively-relevant features of the situations.

In sum, we note that the patterns in data transcend the particulars in which they were gathered, in ways that we can talk about in terms of students' capabilities, which we implement as student model variables and organize in ways tuned to their purpose. Having the latent variables in the student model as the organizing framework allows us to carry out coherent interpretations of evidence from a task with one set of surface features to other tasks that may be quite different on the surface. The machinery of probability-based inference in the evidence accumulation process is used to synthesize information from diverse tasks in the form of evidence about student capabilities, and quantifies the strength of that evidence. Psychometric models can do these things to the extent that the different situations display the pervasive patterns at a more fundamental level, because they reflect fundamental aspects of the ways students think, learn, and interact with the world.

4.2 Is Educational Data Mining Psychometrics?

The preceding section described key ideas of the latent variable models that represent advanced application of psychometrics in educational assessment, primarily under the standard assessment paradigm. How does EDM relate to these ideas?

We stated earlier that there is a broad range of analytic needs in assessment, ranging from support of domain analysis of text, feature extraction of complex logs, and methods for inferring connections among assessment activities [Behrens et al. 2012]. Many of these activities are not addressed by traditional psychometric models. Nor are they meant to be; the majority of psychometric activity under the standard assessment paradigm has focused on t he evidence accumulation process, assuming other processes are sufficiently well prescribed by the multiple-choice and ordered score categories of the standard assessment paradigm.

Using EDM techniques to detect relevant patterns in lower-level raw data is not psychometrics in terms of the key ideas outlined above, but it is it undertaking foundational work for broad psychometric modeling; this is feature generation and feature selection in Figure 4. Such patterns that are grist for defining the observed score variables (i.e., evidence identification, in terms of the four-process delivery system architecture) serve as input to psychometric models. EDM techniques are ways for discovering or iteratively refining data variables from complex performances.

As examples, Kerr and Chung [this issue] conducted exploratory cluster analyses to identify salient features of student performance in an educational video game targeting rational number addition, and Hershkovitz and Nachmias [2010] used *learnograms* to identify variables indicative of student motivation from logs of student activities in an online learning system for Hebrew vocabulary. What is missing from this EDM work from the perspective of psychometrics, though, is the dependence of these variables on the latent variables.

Obtaining summary measures of aspects of students' performance in a particular complex task and taking the scores at face value does not incorporate the probabilistic contribution of psychometrics. There may indeed be valuable information about the performance and about the student, but there is no characterization of the evidentiary value of the evidence or of its meaning outside the framework of the particular task.

We can, however, incorporate the key psychometric idea of quantifying the value of evidence by using replicate tasks or internal measures of variability such as *jackknife standard errors* [Mosteller and Tukey 1977]. For example, Beck [2005] introduced engagement tracing based on response times, and presented reliability evidence for the use of response times to multiple-choice cloze questions based on split-half methods. The machinery is in place to experiment with alternative scoring methods or data capturing procedures to improve the value of the evidence from these particular task situations.

An EDM model that includes latent variables which are posited to account for observable score variables through conditional probability distributions, yet remains bound to a particular task or set of tasks, is very close to the spirit of latent variable psychometrics. The final step is whether the same student model comprising these latent variables can be used to model performance on different tasks in the domain - even ones that appear idiosyncratically in games or simulations, when recognizable by virtue of their salient features as instances of classes of recurring situations.

For example, Arroyo et al. [2010] used BNs with latent variables to model unknown student attitudes and goals (e.g., fear of being wrong, wanting a challenge) in a w eb-based tutoring systems for high school mathematics. The extent to which the interpretations of latent variables representing the student attitudes and goals are restricted to the particular tasks in the system or are generalizable to other high school mathematics tasks - or their attitudes with respect to other academic domains - is unclear. Absent empirical studies, the argument for generalizability of the interpretations of the latent variables rests on one of design, drawing strength from a principled approach to design and the coherence among the domain analysis, domain modeling, CAF, assessment implementation, and assessment delivery layers of the assessment design and implementation process.

4.3 Data Mining as a Reaction to Perceived Limitations of Psychometrics

The evidentiary reasoning insights of psychometrics are quite powerful for familiar kinds of assessments. A century of experience and research and an armamentarium of models and techniques exist for modeling data that consist of item scores and judges’ ratings of performances. Far less guidance is available for modeling the kinds of work that can now be routinely captured in digital environments: every key stroke and time stamp in the log of an open-ended troubleshooting task in a simulation environment, for example, or the real-time interactions of hundreds of players in an online game, or continuous physical monitoring of students as well as their actions, or step-by-step task solutions from thousands of students on hundreds of problems in intelligent tutoring environments as their proficiencies grow over the course of study.

With a few exceptions [see, e.g., Ramsay 1982, on the psychometrics of functions as data] there simply are not many tools on the traditional psychometric shelf to make sense of the complex forms of data that are becoming quite routine. On the whole, however, the historically coarse-grained and sparse nature of assessment data has led to a greater focus on not only these kinds of data, but higher-level psychological constructs. This stands in contrast to more detailed, richer, and interactive data and finer-grained modeling of students' processes and strategies. It is thus natural to adapt machinery from other fields that deal with masses of data, such as physics, biology, meteorology, intelligence analysis, and computational linguistics, to bear on educational problems. We would argue that we can improve assessment practice by integrating concepts and machinery from the psychometric tradition and the EDM tradition, integrated within the broad assessment perspective reflected in ECD.

4.4 Leverage Points for Educational Data Mining

There are three particular leverage points for EDM with respect to psychometric modeling. They concern (1) the modeling of student proficiencies (i.e., the latent variable characterization of aspects of students' capabilities), (2) understanding salient patterns in raw feature data required for evidence identification, and (3) understanding relationships between features of evolving situations and students' proficiency-driven actions within those situations. We can categorize them by the three main models in the CAF.

4.4.1 Student Models. These are the latent variable models, the semantics of which refer to aspects of students' capabilities as they might apply across different situations, in terms that can be applied to model probabilities across different situations. The constituent statistical variables need not look at all like familiar test scores.

Gitomer and Yamamoto's [1991] model for understanding logic gate problems, for example, was a DCM with student model variables for understanding basic operations and common misconceptions that students might hold. The model could be used both for isolated problems and for reasoning within larger simulation tasks. Mislevy and Gitomer's [1996] model for hydraulics troubleshooting was a BN, with variables for troubleshooting strategies and knowledge of subsystems.

The design objective is to discover, develop, and refine student models that are at once consistent with the data, substantively meaningful, and practically useful for the job at hand. What should the variables be? How many, what is their nature, are there relationships among them such as prerequisition or conjunction? Methods from EDM that can be brought to bear on these questions *include self organizing maps* [Pirrone et al. 2003], *association rule mining* [Garcia et al. 2010], *sequential pattern analysis* [Zhou et al. 2010], *and process mining* [Trcka et al. 2010]. There is a clear overlap between such EDM models and psychometric models including factor analysis, latent class analysis, cluster analysis, and BNs as they are used to address this challenge.

4.4.2 Evidence Models. This is perhaps the focus of most interest in mining massive data from complex performances, especially in interactive digital environments. There is not a lot of experience in psychometrics for this kind of data, and it is exactly such data that many EDM techniques have been designed to explore. It is easy to amass rich and voluminous bodies of low-level data, mouse clicks, cursor moves, sense-pad movements, and so on, and choices and actions in simulated environments. Each of these bits of data, however, is bound to the conditions under which it was produced, and does not by itself convey its meaning in any larger sense. We seek relevance to knowledge, skill, strategy, reaction to a situation, or some other situatively and psychologically relevant understanding of the action. We want to be able to identify data patterns that recur across unique situations, as they arise from patterns of thinking or acting that students assemble to act in situations. It is this level of patterns of thinking and acting we want to address in instruction and evaluation, and therefore want to express in terms of student model variables.

The following examples illustrate techniques that assessment researchers have used along with domain theory to discover data patterns that evidence psychological patterns:

1. In troubleshooting, using logic rules to identify action sequences in space-splitting situations as consistent with space-splitting, serial elimination, remove-and-replace, redundant, and irrelevant [Mislevy and Gitomer, 1996].
2. In evaluating speaking skills in a language testing, using supervised neural networks to identify phonemes, then words, in acoustic streams [Bernstein 1999].
3. In marksmanship training, using graphical analysis to identify and correlate patterns of breathing and trigger break timing [Chung et al. 2011].
4. In an epidemiology simulation, using unsupervised neural networks to discover patterns of systematic and haphazard sequencing of tests [Hurst et al. 1997].

We note that it is not the data patterns in and of themselves that matter in assessment, but how data patterns provide evidence of capabilities that are relevant to the purpose of the assessment. Just having gigabytes of keystrokes and mouseclicks is not sufficient for claiming one has good evidence for a particular purpose. In fact, the process of discovering and using data patterns is iterative, in that we capture data (based on c urrent understanding), identify salient higher-level features (that we can use these operationally), and continue mining lower level data and using our insights to improve the design of situations for students to act in and features of their performances to capture and interpret.

For example, as noted earlier, Kerr and Chung [this issue] report on cluster analyses of attempts in an educational video game in which it was found that, in some cases, students successfully completed the levels of the game (i.e., solved tasks) using strategies other than the intended strategy. Importantly, such alternative solution strategies that worked in early levels (easier tasks) were ineffective on later levels (harder tasks). With this finding, the designers of the game (assessment) can reorder or redesign game levels (tasks) so that such strategies do not yield a solution.

4.4.3 Task Models. Although data mining of features of situations is entwined with data mining of features of performance, we break it out separately here because it has been until recently a neglected area in psychometrics. As noted previously, the key point is that students' actions make sense only in terms of the situations they are in. This insight was easy to slide over with standard tests, because tasks were fabricated by expert test developers, who knew what features to build into them to evoke what kinds of evidence of knowledge and skill. It was enough for a psychometrician to know simply that the features were there, and she had only to focus on performance data, say right or wrong answers. The problem of how situation features determine the meaning of performance features cannot be avoided in continuous, evolving, and digitally mediated performance tasks such as in games and simulations.

These performance tasks may include fixed-form work products, such as interim reports and final solutions that can be modeled using modest extensions of familiar psychometric techniques. But the moment-by-moment situations that students act in, and from which the bulk of data may be obtained, arise idiosyncratically from students' performances and the system's responses to them. It is necessary to recognize recurring and substantively salient features of situations, so that salient features of performance in those situations can be recognized and evaluated.

The preceding example of identifying space-splitting situations in hydraulics troubleshooting was of this character in that it was necessary to parse not only the state of the aircraft system but also the information that could be known to the student from his earlier actions. Similarly, examples in language testing are dyads of speech acts in conversations. A historical example is computer chess, where the challenge is to be able to characterize positions in terms of features such as pawn structure and phase of game. An automated evaluation scheme must be able to characterize strength of positions in order to compare them before and after sequences of possible moves. In other words, one can't make sense of a move without jointly making sense of situation [Bleicher et al. 2010].

4.5 Additional Leverage Points for Data Mining in Assessment

In addition to the traditional emphasis on the analysis of work products and the psychometric touch points discussed above, we consider how EDM can provide opportunities to improve assessment practice in earlier layers of the ECD framework, namely the domain analysis and domain modeling layers that have traditionally been viewed as prior to modeling and analysis activity; recall that Table 2 provides a summary of the possible applications discussed above as well as in this section.

4.5.1 Domain Analysis. For the educational data miner, including domain analysis into the assessment framework means that the increasingly available corpora of information available in digital form and related techniques for information extraction and knowledge management from extant text can be used to improve assessment in new ways [Behrens et al. 2012]. They can be used to inform understandings of how ideas are represented and used in practice, to inform not only curriculum and instruction, but to continually shape assessment activity. Consider for example that knowledge management in the social sciences remains largely a manual task for researchers. New techniques of mining scientific publications can help inform scientists about emerging concepts or data in ways that can likewise help assessment developers track such changes.

4.5.2 Domain Modeling. Here again we see a valuable and emerging role for the data miner. Behrens et al. [2012], for example, discuss the emerging use of the data-mining approaches of semantic web technologies (e.g., the *Achievement Standards Network*) to articulate component relationships between standards across different educational systems - typically at the country level - and content associated with those standards or other standards whose relationship can be implied via machine induction over the *Resource Description Framework* space.

This type of hierarchy or relational analysis combined with *natural language processing* (NLP) [Manning and Schütze 1999] possibilities available for use in domain analysis may lead to important insights into the interconnection of different types of concepts, standards and assessment activities.

In previous sections we discussed the role of EDM for improving psychometric aspects of the CAF (i.e., the student and evidence models) with implications for assessment delivery. We think that as the conceptualization of assessment continues to broaden to include the end-to-end operationalization of assessment even more opportunities for EDM in assessment will arise. For example, in the area of task generation, schemes based on new approaches such as crowd sourcing are evolving that will require new tracking and connections of data. Likewise, as reporting of assessment results continues to move to on-line formats, the data available from the use of the reports and inferences about their design and communicative value may become part of the standard domain for educational data analysis and EDM.

5. CONCLUSION

ECD is a comprehensive framework for describing and understanding assessment activities. This paper has discussed the central components and logical features of ECD, highlighting the evidentiary focus in obtaining, interpreting, and explaining data. We discussed the role of generalization and latent variables in assessment thought and how psychometric models support this conceptualization. EDM is an emerging technology that provides important insights in several layers in the ECD framework if applied within an integrated assessment design and implementation endeavor, and can broaden the reach of computational activity to refine and automate assessment.

Key to the understanding of this interplay is that though ECD stresses the importance of a priori clarity in the evidentiary arguments that drive design and development of assessment, this clarity is likewise informed by insights from a posteriori analysis. For example, the development of a networking skill performance assessment system in the CNA included the specification of scoring rules based on detailed expert analysis and student protocol analysis constructed from an ECD framework [Williamson et al. 2004].

However, post hoc analysis of log data using NLP algorithms revealed that the empirical log data was not consistent with the theoretical model implied in the scoring rules [DeMark and Behrens 2004]. The interplay of data and scoring theory is especially important in new domains where there may be an absence of theory concerning the observable representations of expertise to inform evidence model specification. In short, assessment design and development needs BOTH rigorous evidential logic AND data analytic insight to support the overall evidentiary logic.

We maintain that EDM, like all data analysis [Behrens and Smith 1996], is best practiced in concert with, rather than isolated from, a theoretical or substantive layer that involve choices or interpretations made by researchers based on purpose or focus of the assessment. Scholars operating within the EDM tradition have advocated as much in applications such as association rule mining [Garcia et al. 2010], sequential pattern analysis [Zhou et al. 2010], cluster analyses of students [Amershi and Conati 2010] and observations [Hershkovitz and Nachmias 2010], and latent variable modeling with BNs [Pardos et al. 2010].

We have no doubt that some interpretation will come from novel or surprising findings when data are analyzed - in this way, EDM represents a way to realize the illuminative goals of exploratory data analysis [Tukey 1977; Behrens 1997; Behrens et al. in press].

However, we add that assessments will be well served if, as much as possible, interpretative aspects are built in a priori through principled assessment design and a posteriori through empirical results. In sum, good EDM in assessment contexts is best viewed in terms of evidentiary reasoning using the lens of ECD. Such a perspective offers (a) a prescriptive approach for assessment design that builds the validity argument concurrently with the assessment, and (b) a framework for recognizing new findings from data analysis and process for refining assessments.

In the current era of rapid technological change and the resulting dramatic impact on a ssessment, new forms of presentation and work product data are being called upon for use in assessment inference. This sea-change highlights the limitations of fixed-response paradigms, and invites the use of EDM to inform the evidence identification process and feed appropriate information to the psychometric (or deterministic) evidence accumulation processes. New types of log files, interactional data streams, and other rich work products require both psycho-social theory and theory generation based on data analysis.

Finally, we emphasize that EDM should not be limited to either just the outputs of scoring processes or just the work products as the inputs to scoring processes. Evolutions in the understanding and practice of assessment call for a broader range of concepts and method to be applied consistently with the notion of a broader assessment perspective. Assessment design, development, delivery, and maintenance processes are complex and increasingly digital. Continued evolution of EDM techniques for understanding the structure of data that affects assessment design, the tracking of tasks over time and differential performance, variations in the task attributes that may be overlooked by hu man coders, and many other artifacts of assessment are increasingly digital and likely to benefit from the application of EDM techniques.

ACKNOWLEDGMENTS

The work of the first author was supported in part by the Center for Advance Technology in Schools (CATS), PR/Award Number R305C080015, as administered by the Institute of Education Sciences, U.S. Department of Education. The findings and opinions expressed in this report are those of the authors and do not necessarily reflect the positions or policies of the Center for Advance Technology in Schools (CATS), the National Center for Education Research (NCER), the Institute of Education Sciences (IES), or the U.S. Department of Education. John Behrens is an Adjunct Assistant Research Professor in the Department of Psychology at the University of Notre Dame. He would like to thank the Department and University for its support.

REFERENCES

ALMOND, R. G., STEINBERG, L. S., and MISLEVY, R. J. 2001. A sample assessment using the four process framework. CSE Technical Report 543. The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA, Los Angeles, CA. Retrieved from http://www.cse.ucla.edu/products/reports/TECH543.pdf

ALMOND, R.G., STEINBERG, L.S., AND MISLEVY, R.J. 2002. Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment, 5*. Retrieved from http://escholarship.bc.edu/ojs/index.php/jtla/article/viewFile/1671/1509

ALMOND, R.G., STEINBERG, L.S., AND MISLEVY, R.J. 2003. A framework for reusing assessment components. In *New Developments in Psychometrics*, H. YANAI, A. OKADA, K. SHIGEMASU, Y. KANO, AND J.J. MEULMAN, Eds. Springer, Tokyo, Japan, 281-288.

AMERSHI, S., AND CONATI, C. 2010. Automatic recognition of learner types in exploratory learning environments. In *Handbook of Educational Data Mining*, C. ROMERO, S. VENTURA, S. VIOLA, M. PECHENIZKIY, AND R. BAKER, Eds. Chapman and Hall/CRC, Virginia Beach, VA, 213-229.

ARROYO, I., COOPER, D. G., BURLESON, W., AND WOOLF, B. 2010. Bayesian networks and linear regression models of students' goals, moods, and emotions. In *Handbook of Educational Data Mining*, C. ROMERO, S. VENTURA, S. VIOLA, M. PECHENIZKIY, AND R. BAKER, Eds. Chapman and Hall/CRC, Virginia Beach, VA, 323-338.

BECK, J. 2005 Engagement tracing: Using response times to model student disengagement. In *Proceedings of the 12th International Conference on Artificial Intelligence in Education*, IOS Press, Amsterdam, The Netherlands, 88-95.

BEHRENS, J. T. 1997. Principles and procedures of exploratory data analysis. *Psychological Methods, 2*, 131-160.

BEHRENS, J.T., COLLISON, T.A., AND DEMARK, S.F. 2005. The Seven Cs of comprehensive assessment: Lessons learned from 40 million classroom exams in the Cisco Networking Academy Program. In *Online Assessment and Measurement: Case Studies in Higher Education, K-12 and Corporate*, S. HOWELL AND M. HRICKO, Eds. Information Science Publishing, Hershey, PA, 229-245.

BEHRENS, J. T., DICERBO, K. E., YEM, N., LEVY, R. in press. Exploratory data analysis. In *Handbook of Psychology*, 2nd *ed., Volume II: Research Methods in Psychology*, W. F. Velicer and I. Winer Eds. Wiley and Sons, New York, NY.

BEHRENS, J.T., FREZZO, D.C., MISLEVY, R.J., KROOPNICK, M., AND WISE, D. 2008. Structural, Functional and Semiotic Symmetries in Simulation-Based Games and Assessments. In *Assessment of Problem Solving Using Simulations*, E.L. BAKER, J. DICKIESON, W. WULFECK, AND H.F. O'NEIL, Eds. Erlbaum, New York, NY, 59-80.

BEHRENS, J.T., MISLEVY, R.J., DICERBO, K.E., AND LEVY, R. 2012. Evidence centered design for learning and assessment in the digital world. In *Technology-Based Assessments for 21st Century Skills: Theoretical and Practical Implications from Modern Research*, M. MAYRATH, J. CLARKE-MIDURA, AND D. H. ROBINSON, Eds. Information Age Publishing, Charlotte, NC, 13-54.

BEHRENS, J. T., AND SMITH, M. L. 1996. Data and data analysis. In *Handbook of Educational Psychology*, D. C. BERLINER AND R. C. CALFEE, Eds. MacMillan, New York, NY, 945–989.

BEJAR, I.I. 2010. *Can Speech Technology Improve Assessment and Learning? New Capabilities May Facilitate Assessment Innovations*. RDC-15. Princeton: Educational

Testing Service. Available online at http://www.ets.org/research/policy_research_reports/rdc-15 .
BERNSTEIN, J. 1999. *PhonePass Testing: Structure and Construct*. Ordinate Corporation, Menlo Park, CA.
BLEICHER, E., HAWORTH, G. M., AND VAN DER HEIJDEN, H. M. J. F. 2010. Data-mining chess databases. *ICGA Journal*, 33 (4), 212-214.
CHUNG, G. K. W. K., NAGASHIMA, S. O., DELACRUZ, G. C., LEE, J. J., WAINESS, R., AND BAKER, E. L. 2011. *Review of Rifle Marksmanship Training Research*. CRESST Research Report. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. Retrieved from http://www.cse.ucla.edu/products/reports/R783.pdf
CRONBACH, L.J. 1980. Validity on parole: How can we go straight? New directions for testing and measurement: Measuring achievement over a decade. In *Proceedings of the 1979 ETS Invitational Conference*, Jossey-Bass, San Francisco, CA, 99-108.
CRONBACH, L.J. 1988. Five perspectives on validity argument. In *Test Validity*, H. Wainer and H. Braun, Eds. Lawrence Erlbaum, Hillsdale, NJ, 3-17.
CRONBACH, L.J., GLESER, G.C., NANDA, H., AND RAJARATNAM, N. 1972. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. John Wiley, New York, NY.
DE AYALA, R.J. 2009. *The Theory and Practice of Item Response Theory*. Guilford Press, New York, NY.
DEMARK, S. F. AND BEHRENS, J. T. 2004. Using statistical natural language processing for understanding complex responses to free-response tasks. *International Journal of Testing, 4*, 371-390.
DICERBO, K. E. 2007. Knowledge structures of entering networking students and their instructors. *Journal of Information Technology Education, 6*, 263-277. Retrieved from http://www.jite.org/documents/Vol6/JITEv6p263-277DiCerbo252.pdf
DICERBO, K. E. 2009. Communicating with instructors about complex data analysis. Paper presented at the annual meeting of the *American Educational Research Association*, San Diego, CA. Retrieved from https://research.netacad.net/mod/data/view.php?d=1&rid=29
DICERBO, K. E. AND BEHRENS, J. T. 2012. Implications of the digital ocean on current and future assessment. In *Computers and Their Impact on State Assessment: Recent History and Predictions for the Future*, R. LISSITZ AND H. JIAO, Eds. Information Age Publishing, Charlotte, NC, 273-306.
EMBRETSON, S. 1983. Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.
EMBRETSON, S.E. (Ed.) 1985. *Test Design: Developments in Psychology and Psychometrics*. Academic Press, Orlando, FL.
FREZZO, D.C., BEHRENS, J.T., AND MISLEVY, R.J. 2009. Design patterns for learning and assessment: facilitating the introduction of a complex simulation-based learning environment into a community of instructors. *The Journal of Science Education and Technology*. Springer Open Access http://www.springerlink.com/content/566p6g4307405346/
FREZZO, D.C., BEHRENS, J.T, MISLEVY, R.J., WEST, P., AND DICERBO, K.E. 2009. Psychometric and evidentiary approaches to simulation assessment in Packet Tracer software. In ICNS '09: *Proceedings of the Fifth International Conference on Networking and Services*, IEEE Computer Society, Washington, D.C., 555 – 560.
GARCIA, G., ROMERO, C., VENTURA, S, DE CASTRO, C., AND CALDERS, T. 2010. Association rule mining in learning management systems. In *Handbook of*

Educational Data Mining, C. Romero, S. Ventura, S. Viola, M. Pechenizkiy, and R. Baker, Eds. Chapman and Hall/CRC, Virginia Beach, VA, 93-106.

GITOMER, D. H., AND YAMAMOTO, K. 1991. Performance modeling that integrates latent trait and class theory. *Journal of Educational Measurement, 28*, 173–189.

GULLIKSEN, H. 1961. Measurement of learning and mental abilities. *Psychometrika, 26*, 93-107.

HAMBLETON, R., AND SWAMINATHAN, H. 1985. *Item Response Theory: Principles and Applications*. Kluwer Nijhoff Publishing, Boston, MA.

HERSHKOVITZ, A., AND NACHMIAS, R. 2010. Log-based assessment of motivation in online learning. In *Handbook of Educational Data Mining*, C. ROMERO, S. VENTURA, S. VIOLA, M. PECHENIZKIY, AND R. BAKER, Eds. Chapman and Hall/CRC, Virginia Beach, VA, 287-297.

HURST, K., CASILLAS, A., AND STEVENS, R. 1997. *Exploring the Dynamics of Complex Problem-solving with Artificial Neural Network-based Assessment Systems.* (CRESST Report 444). National Center for Research on Evaluation, Standards and Student Testing (CRESST), Los Angeles, CA. Retrieved from http://www.cse.ucla.edu/products/reports/TECH444.pdf

ISELI, M.R., KOENIG, A.D., LEE, J.J., AND WAINESS, R. 2010. *Automatic Assessment of Complex Task Performance in Games and Simulations*. CRESST Technical Report 775. University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles, CA. Retrieved from http://www.cse.ucla.edu/products/reports/R775.pdf

JENSEN, F.V. 1996. *An Introduction to Bayesian Networks*. Springer, New York, NY.

JUNKER, B. 2011. The role of nonparametric analysis in assessment modeling: Then and now. In *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, N. J. Dorans and S. Sinharay Eds. Springer, New York, NY, 67-85.

KANE, M.T. 1992. An argument-based approach to validation. *Psychological Bulletin, 112*, 527-535.

KANE, M.T. 2006. Validation. In *Educational Measurement*, 4th ed., R. L. Brennan, Ed., American Council on Education and Praeger Publishers, Westport, CT, 17-64.

KERR, D., AND CHUNG, G. K. W. K. this issue. Using cluster analysis to identify key features of student performance in educational video games and simulations. *Journal of Educational Data Mining*.

LEIGHTON, J. P., AND GIERL, M. J. (Eds.) 2007. *Cognitive Diagnostic Assessment for Education: Theory and Practices*. Cambridge University Press, Cambridge, UK.

LEVY, R., AND MISLEVY, R.J. 2004. Specifying and refining a measurement model for a simulation-based assessment. *International Journal of Testing, 4*, 333-369.

LEVY, F., AND MURNANE, R.J. 2004. *The New Division of Labor: How Computers are Creating the Next Job Market*. Princeton University Press, Princeton, NJ.

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

LIU, M., AND HAERTEL, G. 2011. *Design Patterns: A Tool to Support Assessment Task Authoring*. Large-Scale Assessment Technical Report 11. Menlo Park, CA: SRI International. Retrieved from http://ecd.sri.com/downloads/ECD_TR11_DP_Supporting_Task_Authoring.pdf

LORD, F.M. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum, Hillsdale, NJ.

MANNING, C.D. AND SCHÜTZE, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

MAYRATH, M., CLARKE-MIDURA, J. AND ROBINSON, D.H. 2012. *Technology-Based Assessments for 21st Century skills: Theoretical and Practical Implications from Modern Research*. Information Age Publishing, Charlotte, NC.

MESSICK, S. 1989. Validity. In *Educational Measurement*, 3rd ed., R. Linn, Ed., American Council on Education, Washington, D.C., 13-103.

MESSICK, S. 1994. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

MISLEVY, R.J. 1994. Evidence and inference in educational assessment. *Psychometrika, 59*, 439-483.

MISLEVY, R.J. 2006. Cognitive psychology and educational assessment. In *Educational Measurement*, 4th ed., R.L. Brennan, Ed. Greenwood, Phoenix, AZ, 257-305.

MISLEVY, R.J. 2011. *Evidence-centered Design for Simulation-based Assessment. Military Medicine*. CRESST Report 800. University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles, CA. Retrieved from http://www.cse.ucla.edu/products/reports/R800.pdf

MISLEVY, R.J., BEHRENS, J.T., BENNETT, R.E., DEMARK, S.F., FREZZO, D.C., LEVY, R., ROBINSON, D. H., RUTSTEIN, D.W., SHUTE, V.J., STANLEY, K., AND WINTERS, F.I. 2010. On the roles of external knowledge representations in assessment design. *The Journal of Technology, Learning, and Assessment, 8*(2). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1621

MISLEVY, R.J. BEHRENS, J.T., DICERBO, K.E., FREZZO, D.C., AND WEST, P. in press. Three things game designers need to know about assessment. In *Assessment in Game-Based Learning: Foundations, Innovations, and Perspectives*, D. IFENTHALER, D. ESERYEL, AND X. GE, Eds. Springer, New York, NY.

MISLEVY, R.J., AND GITOMER, D.H. 1996. The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction, 5*, 253-282.

MISLEVY, R.J. AND RICONSCENTE, M.M. 2005. *Evidence-Centered Assessment Design: Layers, Structures, and Terminology*. PADI Technical Report. Retrieved from http://padi.sri.com/downloads/TR9_ECD.pdf

MISLEVY, R.J., AND RICONSCENTE, M.M. 2006. Evidence-centered assessment design: Layers, concepts, and terminology. In *Handbook of Test Development*, S. Downing and T. Haladyna, Eds. Erlbaum, Mahwah, NJ, 61-90.

MISLEVY, R.J., RICONSCENTE, M.M., AND RUTSTEIN, D.W. 2009. *Design Patterns for Assessing Model Based Reasoning*. PADI-Large Systems Technical Report 6. Menlo Park, CA: SRI International. Retrieved from http://ecd.sri.com/downloads/ECD_TR6_Model-Based_Reasoning.pdf

MISLEVY, R.J., STEINBERG, L.S., AND ALMOND, R.G. 1999. *On the Roles of Task Model Variables in Assessment Design*. CSE Technical Report 500. The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA, Los Angeles, CA. Retrieved from http://www.cse.ucla.edu/products/reports/TECH500.pdf

MISLEVY, R.J., STEINBERG, L.S., AND ALMOND, R.G. 2003. On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-62.

MOSTELLER, F., AND TUKEY, J.W. 1977. *Data Analysis and Regression: A Second course in Statistics*, 1st ed., Addison Wesley, New York, NY.

MURNANE, R.J., SHARKEY, N.S., AND LEVY, F. 2002. A role for the internet in America education? Lessons from the Cisco Networking Academies. In *The Knowledge Economy and Postsecondary Education*, P. Graham and N. Stacey, Eds. National Academy Press , Washington, DC, 127-157.

PARDOS, Z. A., HEFFERNAN, N. T., ANDERSON, B. S., AND HEFFERNAN, C. L. (2010). Using fine-grained skill models to fit student performance with Bayesian networks. In *Handbook of Educational Data Mining*, C. ROMERO, S. VENTURA, S. VIOLA, M. PECHENIZKIY, AND R. BAKER, Eds. Chapman and Hall/CRC, Virginia Beach, VA, 417-426.
Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. Kaufmann, San Mateo, CA.
PELLEGRINO, J., CHUDOWSKY, N., AND GLASER, R. (Eds.). 2001. *Knowing What Students Know: The Science and Design of Educational Assessment*. National Research Council's Committee on the Foundations of Assessment. National Academy Press , Washington, DC.
PIRRONE, R., COSSENTINO, M., PILATO, G., AND RIZZO, R. 2003. Concept maps and course ontology: A multi-level approach to e-learning. In *Proceedings of the 8th AI*IA Workshop on Artificial Intelligence and E-learning*, Pisa, Italy.
RAMSAY, J.O. 1982. When the data are functions. *Psychometrika, 47*, 379-396.
RECKASE, M.D. 2009. *Multidimensional Item Response Theory*. Springer, New York, NY.
ROMERO, C., VENTURA, S., PECHENIZKIY, M., AND BAKER, R.S.J.D. (Eds.). 2011. *Handbook of Educational Data Mining*. CRC Press, Boca Raton, FL.
RUMBAUGH, J., BLAHA, M., PREMERLANI, W., EDDY, F., AND LORENSEN, W. 1991. *Object-Oriented Modeling and Design*. Prentice-Hall, Englewood Cliffs, NJ.
GINSBERG, M. 1987. *Readings in Nonmonotonic Reasoning*. Morgan Kaufmann, Los Altos, CA.
RICONSCENTE, M., MISLEVY, R.J., AND HAMEL, L. 2005. *An Introduction to PADI Task Templates*. PADI Technical Report 3. SRI International, Menlo Park, CA. Retrieved from http://padi.sri.com/downloads/TR3_Templates.pdf
RUPP, A.A. 2002. Feature selection for choosing and assembling measurement models: A building-block-based organization. *International Journal of Testing, 2*, 311-360.
RUPP, A.A., DICERBO, K.E., LEVY, R., BENSON, M., SWEET, S., CRAWFORD, A.V., FAY, D., KUNZE, K. L., CALIÇO, T., AND BEHRENS, J.T. this issue. Putting ECD into practice: The interplay of theory and data in evidence identification and accumulation within a digital learning environment. *Journal of Educational Data Mining*.
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. *The Journal of Technology, Learning, and Assessment, 8*(4). Retrieved from http://escholarship.bc.edu/jtla/vol8/4
RUPP, A.A., TEMPLIN, J., AND HENSON, R.J. (2010). *Diagnostic Measurement: Theory, Methods, and Applications.* Guilford Press, New York, NY.
SHAFFER, D.W. 2006. Epistemic frames for epistemic games. *Computers & Education, 46*, 223-234.
SHUTE, V. J. 2011. Stealth assessment in computer-based games to support learning. In *Computer Games and Instruction*, S. TOBIAS AND J.D. FLETCHER, Eds. Information Age Publishers, Charlotte, NC, 503-523.
STEINBERG, L.S., MISLEVY, R.J., AND ALMOND R.G. 2005. *Portal Assessment Design System for Educational Testing*. U.S. Patent #434350000, August 4, 2005.
TANNENBAUM, A.S. 2003. *Computer Networks*. Pearson Education, Upper Saddle River, NJ.
TATSUOKA, K.K. 1983. Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.
THEODORIDIS, S. AND KOUTROUMBAS, K. 1999. *Pattern Recognition*. Academic Press, San Diego, CA.

Tukey, J.W. 1977. *Exploratory Data Analysis*. Adison Wesley, Reading, MA.

TRCKA, N., PECHENIZKIY, M., AND VAN DER AALST, W. 2010. Process mining from educational data. In *Handbook of Educational Data Mining*, C. ROMERO, S. VENTURA, S. VIOLA, M. PECHENIZKIY, AND R. BAKER, Eds. Chapman and Hall/CRC, Virginia Beach, VA, 123-142.

VAN DER LINDEN, W.J., AND GLAS, C.A.W. (Eds.) 2010. *Elements of Adaptive Testing*. Springer, New York, NY.

VANLEHN, K. 2008. Intelligent tutoring systems for continuous, embedded assessment. In *The Future of Assessment: Shaping Teaching and Learning*, C.A. Dwyer, Ed. Erlbaum, New York, NY, 113-138.

WAINER, H., DORANS, N.J., FLAUGHER, R., GREEN, B.F., AND MISLEVY, R.J. 2000. *Computerized Adaptive Testing: A Primer*. Routledge, New York, NY

WEST, P., RUTSTEIN, D.W., MISLEVY, R.J., LIU, J., LEVY, R., DICERBO, K.E., CRAWFORD, A., CHOI, Y., CHAPPEL, K., AND BEHRENS, J.T. 2010. *A Bayesian Network Approach to Modeling Learning Progressions*. CRESST Research Report. The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA, Los Angeles, CA. Retrieved from http://www.cse.ucla.edu/products/download_report.asp?r=776

WILLIAMSON, D. M., BAUER, M., STEINBERG, L. S., MISLEVY, R. J., BEHRENS, J. T., AND DEMARK, S. 2004. Design rationale for a complex performance assessment. *International Journal of Measurement, 4*, 333–369.

WILLIAMSON, D.M., MISLEVY, R.J., AND BEJAR, I.I. (Eds.). 2006. *Automated Scoring of Complex Performances in Computer Based Testing*. Erlbaum Associates, Mahwah, NJ.

WITTEN, I.H., AND FRANK, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. Morgan Kaufmann, San Francisco, CA.

ZHOU, M., XU, Y., NESBIT, J.C., AND WINNE, P.H. 2010. Sequential pattern analysis of learning logs: Methodology and applications. In *Handbook of Educational Data Mining*, C. ROMERO, S. VENTURA, S. VIOLA, M. PECHENIZKIY, AND R. BAKER, Eds. Chapman and Hall/CRC, Virginia Beach, VA, 107-121.