Putting ECD into Practice: The Interplay of Theory and Data in Evidence Models within a Digital Learning Environment

ANDRÉ A. RUPP
University of Maryland
ROY LEVY
Arizona State University
KRISTEN E. DICERBO
Pearson
SHAUNA J. SWEET
University of Maryland
AARON V. CRAWFORD
Arizona State University
TIAGO CALIÇO
University of Maryland
MARTIN BENSON
Cisco
DEREK FAY
Arizona State University
KATIE L. KUNZE
Arizona State University
ROBERT J. MISLEVY
Educational Testing Service
and
JOHN T. BEHRENS
Pearson

In this paper we describe the development and refinement of *evidence rules* and *measurement models* within the *evidence model* of the *evidence-centered design* (ECD) framework in the context of the *Packet Tracer* digital learning environment of the *Cisco Networking Academy*. Using *Packet Tracer* learners design, configure, and troubleshoot computer networks within an interactive interface. This leads to *product data*, which result from the students' final submitted network configurations, and *process data*, which are log file entries detailing how they got to the final configurations. We discuss how an iterative cycle of empirical analyses and discussions with subject-matter experts is essential for identifying and accumulating evidence about skill profiles of learners and their development. We present results from descriptive, exploratory, and confirmatory diagnostic modeling analyses for both data types, which required bringing to bear a diversity of tools from multivariate statistics, modern psychometrics, and educational data mining. We close the paper with a discussion of the implications of this work for evidence-based argumentation guided by ECD principles within digital learning environments more generally.

Key words: Educational data mining, evidence-centered design, log files, diagnostic classification models, Bayesian networks

---

Authors' addresses: André A. Rupp, Shauna J. Sweet, and Tiago Caliço, Department of Human Development and Quantitative Methodology, University of Maryland, 1230-A Benjamin Building, College Park, 20742, ruppandr@umd.edu, shaunajsweet@gmail.com, and tcalico@umd.edu; Roy Levy, Aaron V. Crawford, Derek Fay, and Katie L. Kunze, School of Social and Family Dynamics, Arizona State University, PO Box 873701, Tempe, Arizona USA 85287-3701, Roy.Levy@asu.edu, Aaron.Crawford@asu.edu, Derek.Fay@asu.edu, and Katie.Kunze@asu.edu; Martin Benson, 589 S East Street, Amherst, MA, 01002, mbenson@cisco.com; John T. Behrens and Kristen E. DiCerbo, 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; Robert J. Mislevy, Research and Development, Educational Testing Service, MS 12-T, Rosedale Road, Princeton, New Jersey USA 08541, rmislevy@ets.org;.

1. INTRODUCTION

This special issue of JEDM is dedicated to the application of *evidence-centered design* (ECD) [e.g. Mislevy et al. 2003; Mislevy et al. 2006; Mislevy et al. this issue] to the design, implementation, and data analysis of innovative digital learning environments that can be used for diagnostic assessment. At a conceptual level, the ECD framework was created to support assessment developers in making explicit the rationales, choices, and consequences reflected in their assessment design.

While ECD can be applied to the development of any kind of assessment where the a priori definition of constructs and associated variables is meaningful, it is particularly suitable to the development of digitally delivered performance-based assessments that are created in the absence of easily delineable test specifications [e.g. Rupp et al. 2010]. It is in these contexts that the number, complexity, and connectedness of decisions that need to be made about the assessment design and results interpretation are most daunting.

1.1 Evidence Models in ECD

As reiterated by Mislevy et al. in this issue, the key components of a validation argument based on ECD practices are (a) domain analysis, (b) domain modeling, (c) the conceptual assessment framework, (d) the assessment implementation, (e) the assessment delivery, and (f) post-assessment activities. While the work that we describe in this paper will touch necessarily on all of these components to some degree, we are most concerned with activities within the conceptual assessment framework.

More specifically, we are concerned with the interplay between theory and data, via continual exchanges between subject-matter experts and statisticians, in order to develop the *evidence rules* and *measurement models* that make up the *evidence model* of the conceptual assessment framework. Evidence rules refer to the means by which we select particular elements from a student's work product and apply scoring rules to obtain observable values for score variables (i.e. *evidence identification*). Measurement models provide us with the method of combining these observables score variables to make inferences about our constructs of interest (i.e. *evidence accumulation*). We refer the reader to Almond et al. [2002] for a more detailed discussion of the relationships of measurement models and score variables within the conceptual assessment framework and its operationalization in the assessment delivery system.

1.2 Practical Implications of Evidence Models

The evidence model is not just a theoretical component within a conceptual framework that exists for design purposes only; in actual assessment delivery it drives the identification and accumulation of empirical evidence about expertise within the digital learning environment. Its specification helps teams of interdisciplinary experts determine how to evaluate which features of the work product provide evidence about student proficiencies, how to apply rules to those work products to obtain observable values of score variables, and how to connect the observable values to student model variables that represent the proficiencies that are of inferential interest. In other words, the evidence models provide the structure and the logic of the processes that take place in an assessment.

In practice, evidence rules and measurement models are often more complex for digital learning environments than for traditional standardized large-scale assessments. For example, for a multiple-choice question a scoring rule can simply be "if the correct option is chosen, then score the answer as correct; otherwise, score it as incorrect". In a digital learning environment, the submitted work products (e.g. configured computer networks, causal maps, reports of experiments) are often more complex and thus require sets of scoring rules that attend to specific features.

For example, in the network engineering environment that is the focus of this paper, a scoring rule could be "if data can move from one PC to another then score this part of the network as correct; otherwise, score it as incorrect." However, it is often less clear how the resulting scores should be weighted and aggregated which is why one, or often multiple, evidence models need to be specified. Moreover, once we focus our attention on the log files of learners, often consisting of hundreds of commands for individual learners, it is much less clear what appropriate scoring rules should be that produce observable score variables or how these should be synthesized.

1.3 Objectives of Paper

In this paper we articulate both the science and the art of specifying and implementing evidence models within a digital learning environment. We present a case study based on *Packet Tracer* (http://www.cisco.com/web/learning/netacad/course_catalog/PacketTracer.html), a simulation and visualization environment designed to support the teaching, learning, and

assessment of *computer network engineering* skills [Frezzo et al. 2010; Frezzo et al. 2009].

Like many digital learning environments, *Packet Tracer* produces two prototypical types of data structures that we alluded to in the previous subsection. *Product data* arise from the application of prespecified scoring rules to submitted work products; these scoring rules are generated and refined by experts and are then automated within the system. *Process data* arise from the moment-to-moment actions that are captured by the computational engine that is running the digital learning environment, which result in complex *log files* of these action sequences.

In this paper, we first employ various potential measurement models using the product data to examine and test alternative rules for evidence identification and accumulation. We then analyze the process data (i.e. the log files) to similarly define evidence rules that can be used to help translate this mass of data into meaningful observable score variables and resulting aggregates.

While doing this, we build two principal arguments. First, we show how a continual information exchange between subject-matter experts and statisticians is indispensable for building coherent evidence-based narratives for product and process data. Second, we demonstrate how data-analytic work can benefit from a coordinated use of tools from areas as diverse as multivariate statistics [e.g. Lattin et al. 2003], psychometrics [e.g. Raykov and Marcoulides 2011], and educational data mining [e.g. Baker and Yacef 2009; Romero et al. 2011]. Put another way, we argue that the technological advances that digital learning environments provide require a similar match in advancements in assessment design and the methodological tools that are brought to bear to analyze their output.

1.4 Organization of Paper

We have divided our paper into three main sections. In the first main section we describe (a) the *Packet Tracer* environment, (b) the task that learners were asked to solve, and (c) the learner samples used in the subsequent analyses. The second main section is dedicated to the structure and analysis of product data using various measurement models. We specifically describe measurement model analyses with both *diagnostic classification models* [e.g. Rupp et al. 2010] and *Bayesian networks* [e.g. Almond et al. in press].

The third main section is dedicated to the structure and analysis of the process data from *Packet Tracer*. We begin by describing the structure of the log files and subsequent tagging efforts. We then describe specifications of evidence identification rules for translating these work products into observable scores and subsequent efforts to accumulate this evidence with tools from multivariate data analysis and educational data mining. We then describe how learner characterizations from the process data analyses can be empirically connected to learner characterizations from the product data analyses.

We close this paper by articulating the lessons that we have learned in this project for principles and practices around specifying and implementing evidence models for complex data structures in digital learning environments in order to construct a coherent and comprehensive validation argument guided by the ECD framework.

2. THE PACKET TRACER ENVIRONMENT AND LEARNER SAMPLES

The *Packet Tracer* environment that we referred to in the previous section is one of the digital learning environments that are provided as part of the *Cisco Networking Academy* (http://www.cisco.com/web/learning/netacad/index.html), 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.

Courses are delivered at high schools, two- and three-year community college and technical schools, and four-year colleges and universities. Since its inception in 1997, the *Networking Academy* has grown to reach a diverse population of about a million students each year in more than 165 countries. Learners participate in in-class work, laboratory activities on real equipment and the *Packet Tracer* simulation tool, standardized large scale assessments with multiple choice and other selected response formats, and digital gaming activities. As part of Cisco's corporate social responsibility program, most instructional materials, including the *Packet Tracer* environment, are made available to the learners in *Networking Academy* classes for free.

2.1 The Packet Tracer Environment

Packet Tracer is a flexible digital platform for designing, administering, and scoring complex tasks in the area of network engineering. Part of the appeal of *Packet Tracer* for learners is the authentic representations of real-life equipment, which include images of their physical shells, their ports, and their interiors, as well as authentic command line structures for interacting with the devices. The *Packet Tracer* environment also allows

instructors to design novel activities and assessments. Instructors can choose to create relatively simple activities (e.g. troubleshooting an already configured network with a minimum number of PCs, routers, and switches) or rather complex assessments (e.g. setting up, configuring, and troubleshooting a network with multiple PCs, routers, and switches where subparts of the network serve different functions), in addition to administering prewritten activities and assessments.

Figure 1 shows a screenshot of the *Packet Tracer* interface with a relatively complex network structure. Each device is identified with an icon and clicking on the icon brings up a window with both a physical representation of the device and simulated configuration interfaces, which is shown in Figure 2. The toolbars on the top, right, and bottom contain a variety of icons that allow learners to simulate and visualize packets of data moving through the network.

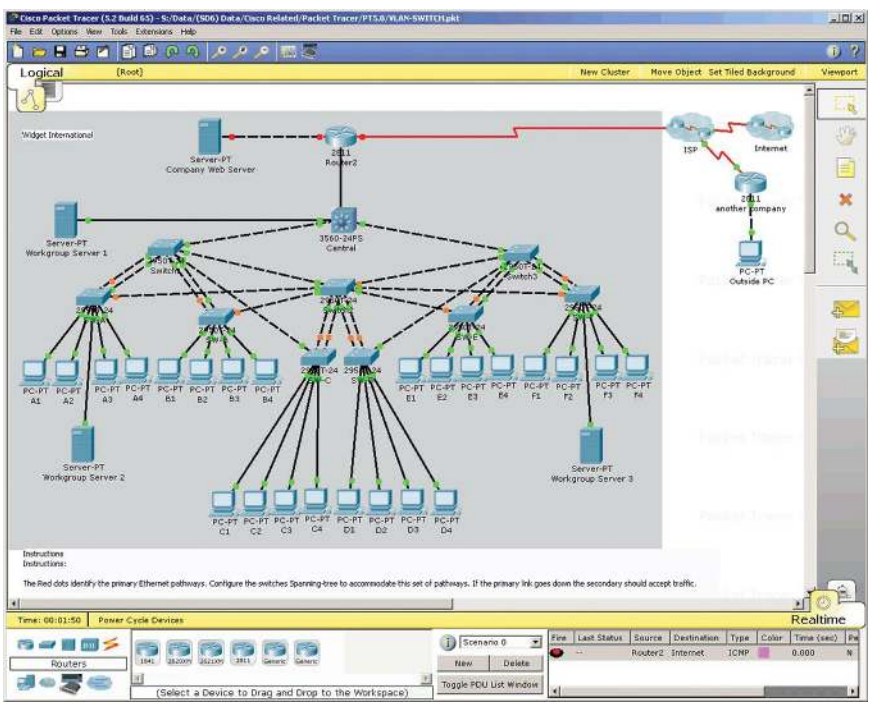

Fig. 1. Screenshot of *Packet Tracer* window with a complex network structure.

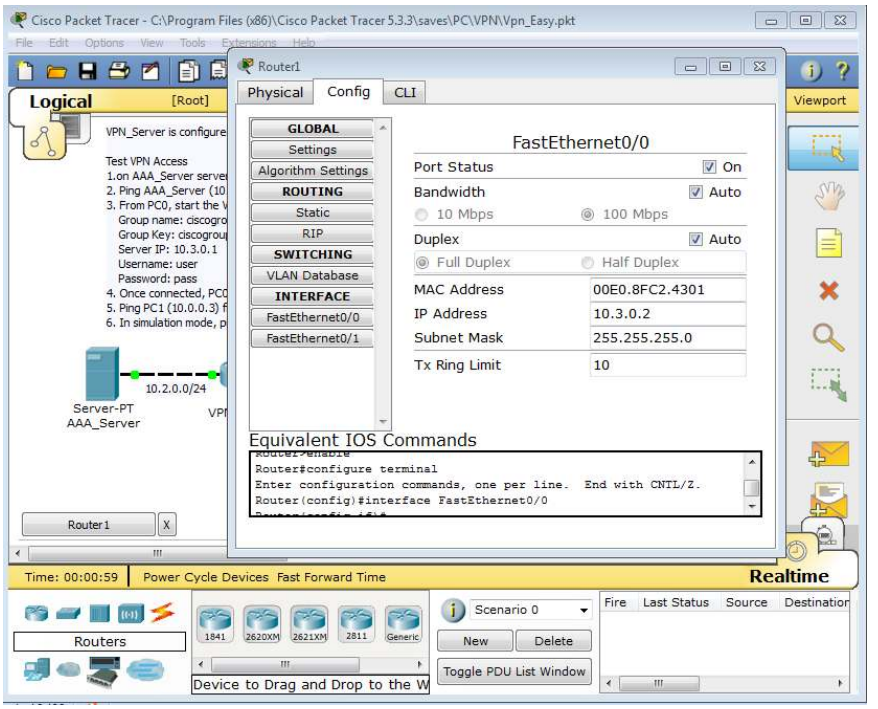

Fig. 2. Screenshot of *Packet Tracer* interface allowing configuration of a router.

The task that is the focus of this paper is a *Packet Tracer Skills-based Assessment* (PTSBA) designed for the *Network Fundamentals* course in the *Exploration* curriculum of the *Networking Academy*. The *Network Fundamentals* course is the first course in a four-course series that teaches introductory networking skills; hence, we refer to this exam as the E1 PTSBA in the following. The E1 PTSBA is most often, although not exclusively, administered as a *low-stakes formative assessment* that prepares learners for the hands-on skills exam that they take at the end of their class in which they demonstrate their learned skills on real equipment.

We note that the network for the E1 PTSBA is much simpler than the network shown in Figure 1 but we cannot show the specific E1 task since it is used in some cases as a summative assessment activity. The E1 PTSBA gives learners a network topology of two PCs, one router, and one switch; the learners are asked to perform basic configuration and troubleshooting tasks to establish communication among these devices. Specifically, learners must assign IP addresses to two router ports and a switch port and troubleshoot a PC IP address. In addition, they must configure the hostnames, banners, and passwords on the router and switch and correctly cable the devices. Further information about the

general design of PTSBAs, including information about the processes of domain analysis and domain modeling within the ECD framework in this context, can be found in Chapple et al. [2009].

2.2 The Learner Samples

The total dataset that was available for the analyses in this paper consisted of 2,269 learners. However, there was some notable heterogeneity in this sample because instructors in the *Networking Academy* use the curriculum and tools provided to them in a variety of ways. For example, instructors can have learners take the E1 PTSBA during class time or can assign it for homework.

Similarly, most instructors have learners take the E1 PTSBA as a formative learning experience, which is the way its use was intended, but some instructors use the assessment for summative grading purposes. It is likely that learners in the latter context are more motivated to perform better than learners in the former context. Thus, we would expect that they typically score somewhat higher and have more effective and efficient solution paths compared to learners who use the E1 PTSBA primarily as a learning tool.

Analysis of data in the online grade book for the purpose of identifying the stakes with which the E1 PTSBA was likely given by instructors indicated that the E1 PTSBA was used for summative high-stakes purposes in approximately 11% of classes. In this data, 280 learners (14.28%) were in this *summative subsample* leaving 1,989 learners in the *formative subsample*. In this paper we do not necessarily always report analyses on both subsamples due to space limitations; however, we do point out places where there were noticeable differences in results. Table I summarizes gender and age characteristics of the combined, formative, and summative subsamples.

Table I. Demographic Breakdown of Learners in Samples

	Combined Sample (n = 2,269)	Formative Subsample (n = 1,989)	Summative Subsample (n = 280)
% Male (% Missing)	83.1 (6.10)	82.10 (7.00)	90.70 (0.00)
Age Mean (SD)	26.94 (8.90)	26.98 (8.99)	26.71 (8.33)
Age Range [Min, Max]	[12.50, 75.30]	[12.50, 70.60]	[15.00, 75.30]

The information shows that learners were predominantly males in their mid- to late-twenties in all three samples, although all three samples ranged from teenagers to septuagenarians.

Table II shows the distribution of geographical regions for the three samples and the reference distributions for the general *Networking Academy* population. The information in Table II shows that the PTSBA exams are used more in some regions than others. For example, over 40% of the sample of PTSBA users is from the United States and Canada, while those countries only make up 14% of *Networking Academy* students. Similarly, only 1% of the sample is from Latin America, while Latin American students make up 18% of the *Networking Academy* population. This is likely because the PTSBAs are currently only offered in English; they are in the process of being translated into the other languages. There were also some regions in which the PTSBAs were not used in a summative manner (i.e. Africa, Greater China, Latin America, and the Middle East).

Table II. Regional Breakdown of Learners in Samples and the *Networking Academy* Population (%)

	Combined Sample	Formative Subsample	Summative Subsample	*Networking Academy*
Africa	2.6	3.0	0.0	5.0
Asia Pacific	7.5	8.2	2.9	19.0
Greater China	8.2	9.4	0.0	7.0
Central and Eastern Europe	7.8	6.8	14.6	7.0
Western Europe	21.8	20.4	32.1	19.0
Latin America & Caribbean	1.0	1.2	0.0	18.0
Middle East	1.4	1.6	0.0	7.0
Russia & CIS	1.9	2.1	0.4	2.0
US and Canada	41.6	40.5	50.0	14.0

Thus, we note some disproportionate representations of learners according to geographical region in our sample data set.

2.3 Evidence Models for Product and Process Data for Packet Tracer

As we noted in the first section, product data in *Packet Tracer* consist of scores that are assigned to the network configurations; specifically, we focus on the final network configuration submitted by the learner at the end of the E1 PTSBA (i.e. the network with all of its connections and device configurations). The software currently evaluates particular features of the work product (e.g. an IP address, a label, or a password), compares them to predefined criteria and rules, and assigns an observable variable value (e.g. "if the submitted IP address is between 192.168.1.1 and 192.168.1.100 it is scored as correct; otherwise, it is scored as incorrect"). All of the current scoring rules lead to binary (i.e. '0' vs. '1') score variables and are fairly straightforward in the context of the E1 PTSBA even though they are more complex for other PTSBAs within the *Networking Academy*.

The E1 PTSBA that is currently used operationally also includes a series of automated evidence accumulation rules that combine the observable score variables into subskill scores and a single total score. The scoring rules are conjunctive / non-compensatory in nature for the subskill levels and compensatory in nature for the total score level; all scoring rules were developed by the subject matter experts. That is, all of the binary observed score variables need to take on a value of '1' (i.e. all of the steps in a subtask need to be solved correctly) for learners to obtain the full credit for the subskill variable. In contrast, total scores were simply computed as the sum of the subskill score variables so that a weaker performance for one subtask could be compensated for by a stronger performance on another subtask. Formal measurement models with latent variable models for testing the empirical viability of the evidence accumulation rules had so far not been applied to the data and are the focus of the second section of this paper.

As we noted earlier, process data consist of log file entries, which are the time-stamped commands learners enter on the devices of the network while they are designing, configuring, and checking it. Other actions within the network that do not lead to command line entries such as click sequences in the GUI are currently not tracked even though such tracking capabilities are currently in development. At the beginning of this project there were no agreed-upon operational scoring rules for the log files even though preliminary evidence identification rules for tagging the log files according to key sets of actions (e.g. management, configuration, and verification actions) had been postulated and tested offline. The work presented in the third section of this paper thus presents the

first serious empirical foray into specifying and empirically utilizing alternative evidence identification and accumulation rules.

2.4 Interplay between Interdisciplinary Expert Teams

As we will show in the remaining sections of this paper, developing evidence rules and measurement models is not a problem that can be solved by statisticians alone. For example, choices need to be made by analysts about the grain size at which both product and process data should be analyzed for a given purpose. The decisions about which variables are important for a particular analysis are made based on an understanding of how observable score variables provide empirical evidence about unobservable proficiency variables in a student model for the reporting and decision-making purpose that is identified.

For example, providing a summative report on a few relatively coarse-grained subskills (e.g. device connection, device configuration, IP addressing, troubleshooting) upon completion of a single or even multiple tasks requires different choices for evidence identification and accumulation than providing fine-grained diagnostic feedback for the purpose of misconception / error analysis and targeted scaffolding either during or after the completion of a single task.

Similarly, indices from measurement models that speak to model-data fit or the degree to which numerical patterns in the output are robust across multiple samples do not provide insight into whether the produced accumulations of evidence are meaningful for the stakeholders who need to interpret them. Resolving issues surrounding the use of assessment information requires a continual and in-depth collaboration between statisticians, content and curriculum developers, learning scientists, assessment experts, and ideally, teachers. This leads to a highly iterative process amongst these collaborators in practice. In other words, as our title for this paper implies, it is neither nuanced theory nor sophisticated statistical analysis alone or separate, but rather their interplay that creates evidentiary richness that carries practical utility for real-life decision making with it.

3. EVIDENCE IDENTIFICATION AND ACCUMULATION FOR PRODUCT DATA

As stated above, PTSBAs in the *Networking Academy* contain evidence identification rules for the product data that automatically score work product features as well as

automated evidence accumulation rules based on the assessment authors' student model for the constellation of knowledge, skills, and abilities assessed.

When following ECD assessment design, the structure of the measurement model that is responsible for evidence accumulation (i.e. that specifies which observable score variables are connected to which unobservable student model variables) is built into the structure of the tasks in that elements of a task are designed to provide evidence about a specific student model variable. In practice, the conceptual measurement models require empirical investigation to validate whether these hypothesized relationships are reflected in data from actual learners. In this section we describe the use of two diagnostic measurement approaches to empirically investigate the fit of the hypothesized measurement model for the product data to the score relationships observed in data.

In describing our modeling approaches we want to underscore three critical methodological points. The first one is that an empirical evaluation of the tenability of a particular combination of observable variables should not only be consistent across descriptive analyses and a single model analysis but, rather, across multiple analyses with different structurally commensurate models. The second point is that different modeling approaches provide, themselves, different empirical 'lenses' onto the issue of combining observable score variables, and that one inevitably learns some unique aspects about this issue from each modeling approach. The third point is that any evaluations of results from statistical models can only be meaningfully made with reference to future decision-making processes through exchanges between the subject-matter experts and the statisticians.

We will specifically describe three modeling approaches that we utilized for our analyses. The first modeling approach that we describe is based on *unidimensional item response theory* (IRT) models [e.g. de Ayala 2009] and associated *classical test theory* (CTT) statistics [e.g. Crocker and Algina 1986], which we applied to observable score variables that were designed to provide coarser summaries of learner performance.

The second modeling approach that we describe is based on a set of statistical models that are known as *cognitive diagnosis / diagnostic classification models* (DCMs) [e.g. Rupp and Templin 2008; Rupp et al. 2010], which we applied to observable score variables that were designed to provide finer summaries of learner performance. They are *restricted latent class models* that assign learners to a particular latent class (i.e. unobserved group) that is associated with a particular skill profile; the models we used

employed a simple 'mastery' vs. 'nonmastery' indicator for each skill. For example, a learner with the profile [1, 0, 1, 0] on four skills is said to have mastered the first and third skills but neither the second nor the fourth skill.

The third modeling approach that we describe is based on *Bayesian networks*, or *Bayes nets*, for short [e.g. Almond et al. in press; Levy and Mislevy 2004], which we also applied to observable score variables; they were constructed to provide finer summaries of learner performance. Bayes nets are conceptually very close to DCMs in terms of modeling objectives and can produce almost identical results in some cases, depending on the model specification and estimation approach. An interesting practical reason for the joint use of DCMs and Bayes nets in this project was the differential expertise of different team members in these areas, which actually allowed the team to explore the theoretical and empirical relationships between these frameworks in ways that are not typically articulated or explored in the applied measurement literature.

We emphasize specifically that we do not seek to describe one 'correct' statistical model. Rather, we demonstrate the complex decision-making processes that are involved in determining multiple 'possible' methodological approaches for determining multiple 'possible' combinations of observable score variables. We show that a combination of expert insight and statistical information can exclude some postulated combinations – as they are either not interpretable, not statistically robust, or neither – but that they do not necessarily directly point to a single preferable scoring solution either. In addition, different methodological approaches have different pros and cons and offer different insights.

3.1 Investigating and Refining an Operational Scoring Structure

Figure A1 shows the relationships among the observable score variables from the E1 PTSBA that is currently used for operational reporting purposes; it also includes a specification of the expert-determined weights that were given to different subsets of variables. We will refer to this scoring structure as the *component-based scoring structure* for the remainder of the paper.

The first level of the scoring structure is formed by 36 so-called *primary observables* that capture the presence or absence of key aspects of the final network configuration at a relatively fine grain size. Consequently, these primary observables are binary score variables that take a value of '1' if a particular aspect of the network is set up and configured properly, and a value of '0' otherwise, as specified in the evidence rules.

The second level of the scoring structure is formed by aggregating the primary observables into a set of 20 relatively coarse-grained *compound observables*. The scoring for the compound variables follows a *conjunctive condensation rule* [e.g. Maris 1999] in that a score of '1' is given on the compound observables only if all of the constituent primary observables are '1', and '0' otherwise. Table A-I shows the names of the primary and compound observables along with a variety of descriptive statistics whose pattern we discuss in the following subsections.

The compound observables are then further aggregated up into four proficiency variables, which are called *performance components* by the development team. These variables represent the primary skill sets that are targeted by the E1 PTSBA and are labeled *device connection* (DC), *basic device configuration* (BDC), *IP addressing* (IP), and *verification and troubleshooting* (VT).

As noted above, the binary scores for the compound observables were weighted by an expert panel based on considerations about the relative importance of each of the components of the final network configuration or necessary actions that are represented by the compound observables. The 20 weighted compound observables add up to 100 points for ease of interpretation, which is a final score aggregation that represents a *global proficiency score* for this assessment.

The statistical aggregations of the primary observables into compound observables, compound observables into performance components, and performance components into a proficiency score, represent a theory-driven process for evidence accumulation. In order to investigate the empirical viability of the particular evidence accumulation that this scoring structure represents, we performed a variety of descriptive, exploratory, and confirmatory analyses.

3.1.1 Descriptive Analyses and Unidimensional IRT Models. Due to space limitations we do not provide a comprehensive summary of all of our descriptive analyses at a fine level of detail here and instead focus on the key messages as they relate to our assessment of the proposed component-based scoring structure described above.

First, we inspected the descriptive statistics for the primary observables, which are based on concepts and formulas in CTT. Specifically, we computed the *proportions correct / p values* as estimates of difficulty as well as the *point-biserial correlation coefficient* and *D index* as estimates of discriminatory power. Following Kline [2005, chapter 5] we computed the *D* index as the difference in *p* values between the lowest- and

highest-scoring 27% of learners. We computed the point-biserial correlation coefficients twice, once using the global proficiency total score and once using the four performance component scores; Table A-I shows these statistics.

Consistent with the hypothesis that the learners in the summative subsample were motivated to perform better than those in the formative subsample, the *p* values for the summative subsample were higher than the corresponding values for the formative subsample. Similarly, with respect to both the single proficiency score and the performance component subscores, the values of the discrimination indices tended to be larger for the formative subsample than the summative subsample, particularly for the *D* index.

Coefficient α estimates of scale homogeneity showed high internal consistency for the global proficiency score for all three samples ($\hat{\alpha}$ = .95 for each). Thus, we estimated the one-, two-, and three-parameter unidimensional IRT models for each of the three samples. We compared both their relative fit – via likelihood-ratio tests and information indices – as well as their absolute fit – via item-fit statistics. In short, the two-parameter IRT model, which allows for differences in difficulty and discrimination across the primary observables, was the best-fitting model in relative and absolute terms across all samples using all of the tests and indices. The estimated reliabilities were .95 for the combined sample, .95 for the formative subsample, and .92 for the summative subsample, which aligned with the Coefficient α estimates of .95 reported in the previous paragraph.

Therefore, the descriptive analyses for the unidimensional proficiency score reflected that this was a relatively easy exam, especially for learners in the summative subsample, which the subject-matter experts viewed as being in alignment with their expectations. This was also reflected in the negatively skewed distributions of the observed global proficiency scores and the estimated latent trait scores from the IRT models in all samples. However, the statistical properties of the primary observables, as captured by CTT and IRT statistics, underscored that the statistical information provided these score variables varied quite a bit.

In contrast, the Coefficient α estimates for the four subscales associated with the DC, BDC, IP, and VT performance components deteriorated somewhat as one would expect when there are fewer pieces of statistical information available per dimension. They specifically ranged from .71 to .93 for the combined sample, from .66 to .93 for the

summative subsample, and from .72 to .93 for the formative subsample. Moreover, the primary observables had similarly mixed measurement properties on these subscales.

3.1.2 Diagnostic Modeling with DCMs. Based on the findings reported above, we decided to use DCMs and Bayes to investigate whether four discrete subscales could be used to accumulate evidence for the four subskills of DC, BDC, IP, and VT. We chose this approach over a traditional approach in, say, multidimensional item response theory [e.g. Reckase 2009] or multidimensional factor analysis [e.g. Thissen and Wainer 2001] because the relatively few primary observables with high discriminatory power that were available for each performance component made a finer differentiation of proficiency levels with continuous subscales statistically challenging, if not impossible.

We estimated the DCMs for both the formative and summative subsample. Due to space limitations, all results in this subsection are for the larger formative subsample only; the patterns for the summative subsample were very similar. The classification using the four subskills of DC, BDC, IP, and VT showed a highly certain classification of the learners into the $2^4 = 16$ possible latent classes; these classes represent all the possible combinations of mastery and nonmastery of the four performance components. That is, the posterior probabilities of membership in one of the 16 latent classes – the one that the learner would be assigned to – exceeded 90% for almost all learners.

In order to criticize the fit of this model to the data, we chose to inspect the *root mean-squared error of approximation* (RMSEA) index for each score variable. Its behavior was relatively well understood from previous simulation work [e.g. Kunina-Habenicht et al. 2012], which suggested a desirable range of values between .00 and about .06 for 'acceptable' model-data fit at the score variable level. Unfortunately, the RMSEA fit index values for the primary observables in this model gave some cause for concern as more than 25% of score variables had values that were larger than desired (min = .02, P25 = .04, median = .07, P75 = .09, max = .13).

Thus, in order to see whether the local model-data fit could be improved, we sought to simplify the model by imposing a restriction on the latent class space that allowed learners to be classified only into a subset of the original 16 latent classes; this subset needed to be specified a priori. We inspected the latent class membership probabilities across the 16 latent classes to identify those latent classes that had either no learners or very few learners in them relative to the remaining latent classes.

Furthermore, feedback from the subject-matter experts was sought to decide whether the skill profiles for the remaining latent classes that we identified as plausible were (a) theoretically defensible and (b) practically useful from an instructional / reporting viewpoint. After a consensus had been reached on this point, we imposed a set of restrictions in the original DCM that reduced the total number of estimated latent classes to five by means of a so-called *attribute hierarchy* [e.g. Tatsuoka 2009]. This improved the distribution of the RMSEA fit index (min = .01, P25 = .03, median = .04, P75 = .05, max = .09) while simultaneously improving classification certainty as the classification problem had now become statistically simpler.

Due to internal coding, the four subskills that define the attribute profile for each latent class were, in sequence, DC, IP, BDC, and VT. Under this restricted DCM, 4.1% of learners were classified as not having mastered any of the four performance components (i.e. they had attribute profile [0, 0, 0, 0]), 18.7% were classified as having mastered only the DC component (i.e., they had attribute profile [1, 0, 0, 0]), 12.2% were classified as having mastered only the DC and BDC components (i.e. they had attribute profile [1, 0, 1, 0]), 11.5% were classified as having also mastered the IP component (i.e. they had attribute profile [1, 1, 1, 0]), and 53.6% were classified as having mastered all four performance components, including VT (i.e. they had attribute profile [1, 1, 1, 1]).

This model structure reflects essentially an ordinal five-class solution that separates learners into five proficiency groups. As shown in Figure 3, the relationship between the latent-class membership variable and the observed total score from CTT as well as the estimated latent trait score (θ) from the two-parameter IRT model were rather strong (Spearman's r = .844, p <.001 for the CTT score and Spearman's r = .852, p < .001 for the IRT score).

As Figure 3 suggests, a one-way ANOVA with latent class membership as the independent variable and post-hoc polynomial contrasts confirmed that the mean total score from CTT follows a cubic trend across the latent classes ($t(1)$ = 6.443, p < .001, partial η^2 = .811) while the mean IRT score follows a linear trend across the latent classes ($t(1)$ = 2.121, p < .001, partial η^2 = .683) in the E1 PTSBA learner population. The subject-matter experts verified that these distributional patterns were, indeed, consistent with the relative easiness of the PTSBA E1 assessment as well as the instructional ordering and cognitive complexity of the four skills.

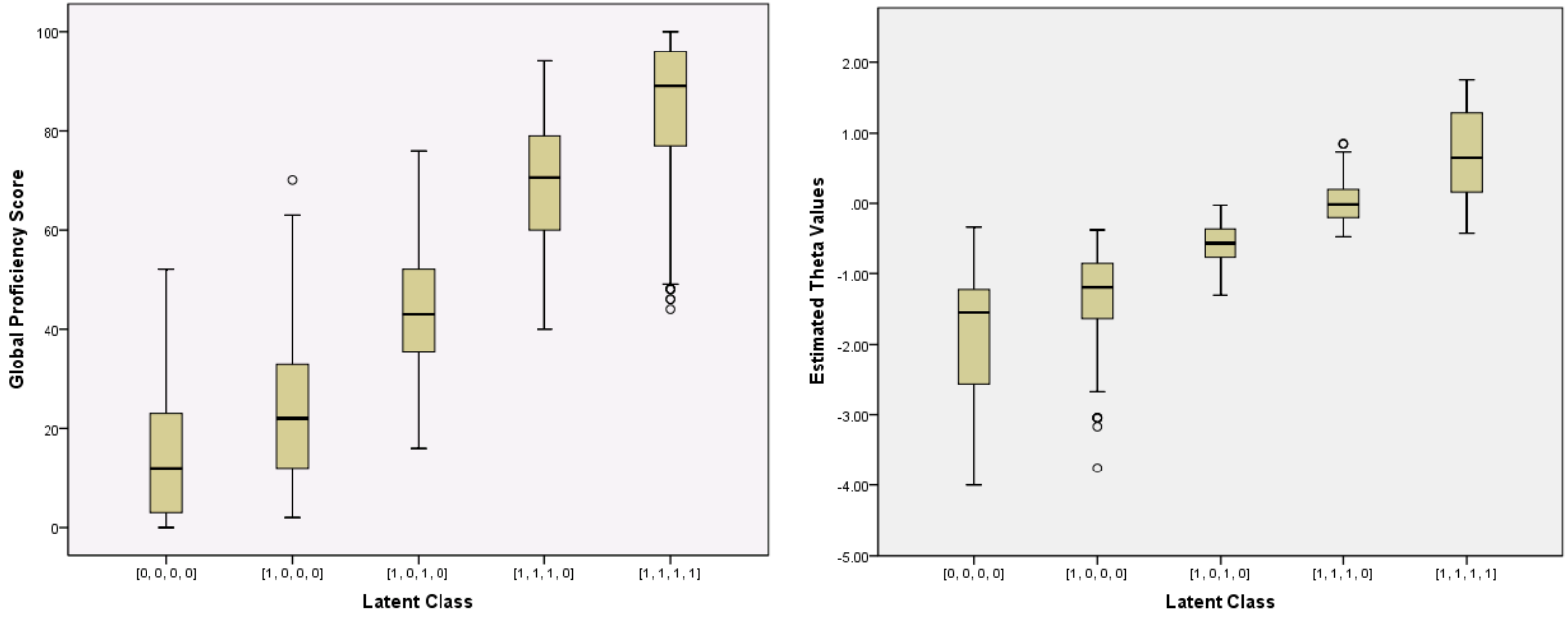

Fig. 3. Proficiency score distributions for DCM analysis of components-based scoring structure with linear hierarchy

3.1.3 Diagnostic Modeling with Bayes Nets. To push the diagnostic analyses one step further, we decided to also estimate a joint measurement model for all three subskill levels of the variable hierarchy (i.e. primary observables, compound observables, and performance components); we used latent variables to represent the compound observables and performance components and used the primary observables as the observable score variables. This could not be done easily within the model specification and estimation framework for DCMs. However, a relatively effective alternative way of specifying, estimating, and refining such a model is with Bayesian nets.

A Bayesian approach to model specification and estimation for the Bayes nets allowed us to rely on a series of very flexible model criticism tools that are often subsumed under the term *posterior predictive model checking* (PPMC) [e.g. Levy et al. 2011; Sinharay 2006]. PPMC proceeds by employing functions called *discrepancy measures* that either capture features of the data or the discrepancy between the data and the model, which one can essentially think of as fit statistics.

Values for the discrepancy measures are computed from the observed data and then compared to values obtained using replicate data drawn from the *posterior predictive distribution*, which represent the expectation of score patterns under the model. For example, PPMC using first- and second-order moments, residualized quantities, and frequencies can help one judge whether the model is able to reproduce the means of individual primary observables, pair-wise associations between pairs of primary observables, or observed score vectors.

Determining a set of the most effective discrepancy measures for Bayes nets for the kinds of data structures we had at hand is an issue of our ongoing research; nevertheless, we utilized a few commonly used measures. Specifically, we inspected three discrepancy measures that reflect different statistical levels of fit, (a) univariate proportions correct / p values for each primary observable, (b) bivariate residuals for pairs of primary observables based on the Q3 statistic [e.g. Yen 1984], and (c) a marginal *generalized dimensional discrepancy measure* (GDDM) [Levy and Svetina 2011] for all of the primary observables as well as for each of the subsets of primary observables as defined by the four performance components.

As noted above, we specified the primary observables as observable score variables – which they are by definition – but specified the compound variables, performance

components, and proficiency variables as higher-order latent variables; Figure A2 shows the structure of this Bayes net, which mimics the scoring structure shown in Figure A1 referenced earlier. Differences between the DCMs and the Bayes net are (a) the lack of compound observables associated with only one primary observable in the Bayes net, for statistical identification reasons, and (b) a saturated structure associating the four performance components in the Bayes net. While this Bayes net was able to reproduce the proportions correct with reasonable accuracy for all primary observables for both learner subsamples, it showed superior model-data fit in terms of the Q3 and GDDM statistics for the summative subsample (see Levy et al. [2011] for additional detailed results using the combined sample).

To demonstrate model-data fit assessment strategies, Figure A3 shows a so-called *heat map* of the *posterior predictive p values* for the Q3 statistic for all pairs of primary observables. One would like to see all, or at least most, of these *p* values close to .50, which would indicate that the observed values of the Q3 statistic are typical of predicted values of the Q3 statistic under the fitted model. In contrast, *p* values that are either rather small or rather large, which are depicted in the figure by white and black squares, respectively, indicate that the observed values are largely atypical and, thus, that the Bayes net does not reproduce well the associations among the observables.

In Figure A3, the formative subsample has more white and black squares than the summative subsample indicating that the Q3 values are better reproduced by the summative subsample. With respect to the GDDM the summative subsample yielded acceptable patterns for the DC as well as VT components while the formative subsample only yielded an acceptable pattern for the DC component.

In order to refine this model, we decided to perform a series of statistical modifications in consultation with the subject-matter experts similar to what we did for the DCM refinement. However, due to the different nature of the discrepancy measures used for the Bayes net compared to the RMSEA statistic for the DCMs this set of modifications took a different path. We used information contained in the Q3 heat map for the summative subsample as a starting point as that model had been most stable to begin with. For each pair of observables that was depicted as a white square in the heat map, the Q3 statistic suggested that the model did not sufficiently represent the observed relationship in the data. In other words, those variable pairs were more strongly related to each other in the data than in the model.

The white squares can therefore be thought of as variable pairs which, according to the data, ought to be more directly connected to each other in the model diagram. Looking at the summative subsample heat map in Figure A3, there were 89 white squares. Increasing the model-implied covariation for even one variable pair could be attempted in a variety of ways, and for each structural change to the model, the model-implied relationships among many variables could be affected. Content experts are in the best position to interpret what the patterns of data-model misfit mean substantively, and are best suited for devising modifications that do not merely capitalize on sampling variability. Our approach was to focus on the areas with the mostly white squares, and to make step-wise modifications to the model structure in consultation with content experts. Improvements to the most problematic areas often improved fit elsewhere as well.

A theme which emerged in this process was that many of the targeted variable pairs consisted of similar tasks on different devices; for example, setting a password on a switch and setting a password on a router. The scoring system had been structured to emphasize the relationships among tasks on a given device. The empirical feedback suggested that in addition to their relationships with other tasks on the same device, some tasks showed residual covariation with similar tasks on different devices. The modifications that were made represented these additional associations among groups of primary observables by either collapsing compound observables or introducing new compound observables.

The model that resulted from this series of modifications showed considerably better fit to the data according to the Q3 statistic as the heat map now contained almost no more white squares. Just as importantly, the model modifications were consistent with domain knowledge. One content expert commented that the modifications represented ideas that had been considered in the initial discussions of the scoring design, but had been discarded because the experts felt constrained to choose between an organization based on device or task.

Empirical evidence thus helped to reintroduce associations between observables that had been contemplated all along but had then been judged to be secondary in importance. The experts who created the scoring system understood that the observables were related in multiple ways, but had no way of quantifying the implications of various choices they had to make. The empirical feedback from the Bayes net allowed them the opportunity to

reevaluate the tradeoffs of increasing the complexity of the scoring model to achieve better fit to the data.

3.1.4. Section Summary. In sum, the three sets of analyses described in this section (i.e. descriptive statistics, diagnostic modeling with DCMs, and diagnostic modeling with Bayes nets) for the component-based scoring structure provided different types of evidence about the measurement model for the E1 PTSBA exam. Descriptive statistics were only useful for detecting statistical issues with individual score variables and dimensional subscores but were not insightful for refining the postulated scoring structure, which is where diagnostic modeling approaches became powerful. The initial DCM and Bayes net analyses for both subsamples provided mixed evidence that the observable score variable combinations that the component-based scoring structure reflected was tenable. However, using model-data fit evaluation / model criticism tools, in conjunction with expert feedback, we were able to determine model structures that showed much better fit to the data.

In the end, both DCM and Bayes-net driven revisions of the component-based scoring structure were statistically defensible even though the model structure for the final Bayes net was more complex than the model structure for the final DCM. Interestingly, from the perspective of classifying learners, there was a very strong alignment between the latent class membership distributions between the final Bayes net and the final DCM. Out of the $2^4 = 16$ latent classes that could be formed by the four higher-level performance component variables in the Bayes net, only nine latent classes had learners in them. In other words, even though the association structure in the Bayes net became more complex, the resulting latent class structure was simplified in terms of actual memberships.

Overall, a total of 86.43% of the learners were classified into latent classes with identical attribute profiles for the DCM and the Bayes net. Thus, integrating our insights from the DCM and Bayes net analyses, we were reasonably confident that a five-class solution was defensible based on the evidence within and across modeling frameworks and the feedback from our experts.

3.2 Procedural Carry-over Effects: Defining a New Scoring Structure

An interesting practical outcome of the continual interchange between the subject-matter experts and statisticians in this project was that the subject-matter experts became

inspired to think differently about evidence identification and accumulation for the E1 PTSBA. In fact, working within the constant feedback loop between statistical information and expert insight made everyone become more aware of the possibilities of modeling different kinds of relationships between observable score variables, rather than thinking in predefined categories such as task boundaries or variable label boundaries. As a result, the team postulated an alternative scoring structure and subsequently investigated it empirically.

The source for developing the alternative scoring structure was a newly refined task bank of traditional standardized assessment tasks whose design specifications had been guided explicitly by ECD principles. What this meant for our data analyses was that the development team had articulated a series of hierarchically organized *claims* about learner proficiency, in ECD parlance. Most importantly, the claims cut across steps that learners had to do in order to design, configure, and troubleshoot the network in the E1 PTSBA; they arguably reflected a slightly stronger, or perhaps more direct, cognitive lens on complex task performance. In some ways, this alignment was already reflected in the results from the model refinement steps from the Bayes net analysis above, but the claims-based framing linked such connections more explicitly to larger assessment efforts and developmental learning progressions of skill sets.

We do not report on the results for these analyses in detail since they follow a similar methodological logic as the descriptions in the preceding subsection; for illustration we show the loading structure and basic CTT statistics in Table A-II and the resulting Bayes net structure in Figure A4. We briefly note, however, that the claims-based DCM showed the best fit amongst all DCMs that we fitted across both scoring structures for both subsamples, which was true both in terms of relative information-based fit indices such as AIC and BIC and the distribution of RMSEA values. The classification certainty was again relatively high for most learners. For example, for the formative subsample, learners were classified into all of the $2^3 = 8$ classes, with only 8.3% of learners being classified as nonmasters of all three claims-based proficiencies and 31.9% of learners being classified as masters of all three proficiencies in alignment with the expected difficulty of the assessment.

The Bayes net analyses showed a more mixed picture of model-data fit. For instance, percentage correct values were again recovered rather accurately for formative and summative subsamples but some, albeit fewer, problems persisted at the pair-wise and

dimensional levels for both subsamples. Specifically, when it came to the GDDMs, neither subsample showed adequate fit for any of the groupings of observables defined by the three claims variables although the summative subsample still showed better fit than the formative subsample.

Therefore, when we connect the process data analyses in the next section of the paper with the product data analyses in this section, we will only use the five-class membership assignments based on the revised DCM and Bayes net for the component-based scoring structure, rather than an assignment based on either DCMs or Bayes nets for the claims-based scoring structure.

4. EVIDENCE IDENTIFICATION AND ACCUMULATION FOR PROCESS DATA

In the previous section we detailed the investigation of measurement models using product data. The use of these measurement models was made possible because the evidence rules for identifying and evaluating features of the final submitted network configurations were already well-specified. Such rules were not specified for process data, however, and the assessment developers did not have explicit hypotheses about what exactly these rules should look like.

In this part of the paper, we thus discuss how a continual exchange between subject-matter experts and statisticians served to identify potential evidence rules and a measurement model for process data. Below we explore the use of summary statistics for log files as observable variables and create a measurement model using *principal component analysis* (PCA) [e.g. Jolliffe 2010]. The PCA analysis allowed us to create two summary dimensions of performance, which can be interpreted as characterizations of the *effectiveness* and *efficiency* of the solution process. We also show how we can connect the scores on these dimensions to the five-class structure from the revised DCM and Bayes net for the product data to arrive at a more integrated, and evidentiary coherent, characterization of learners' performance on the E1 PTSBA.

As noted in the first main section of this paper, the *Packet Tracer* engine allows for an automated tracing that results in complex data in the form of log files. These data now need to be converted into evidence, which is a nontrivial endeavor. While the idea of capturing problem-solving behavior during task completion is certainly not new (e.g. the use of think-aloud protocols as discussed in Leighton [2004]), the unobtrusiveness with which this information about actions can be collected from digital interactions certainly

opens new analytic possibilities. Log files seemingly solve the intrusiveness problem of think-aloud protocols that require learners to interrupt their cognitive response process and rely on a rather sophisticated level of meta-cognition in situ.

However, log files suffer from the opposite problem in that they are raw pieces of data without any direct guidance by the learner as to how they should be interpreted. On the one hand, the fine grain size at which log files typically trace activities of learners allows for a potentially fluid window into the minds of learners while they solve complex tasks. On the other hand, this grain size makes the task of creating evidence rules without a clear link between this large number of actions and sharply aligned theories of cognitive response processes daunting. Contrary to the more confirmatory nature of evidence model work for the analysis of product data that we have described in the previous section, evidence model work for process data thus typically has a more exploratory nature.

Statistically it is thus not surprising that techniques in areas such as *educational data mining* [e.g. Baker and Yacef 2009; Romero et al. 2010] or *business process mining* [e.g. van der Aalst 2011] are more attractive to solving evidence model-related problems for log file data than IRT models, DCMs, or Bayes nets. Of course, none of the resulting output is interpretable without at least a reasonably sound articulation of a theory of how task features and learner characteristics combine to produce classes of activity traces for different learner groups, which is where a framework like ECD is invaluable. As a consequence, just as with product data analyses above, it is crucial to have an interdisciplinary team of subject-matter experts and statisticians working together to arrive at meaningful interpretations for process data analyses.

4.1 Preprocessing / Tagging of Log files

The structure of log files in digital learning environments is often such that a direct analysis of the raw files is practically not meaningful as too many functionally necessary, but evidentiary meaningless, aspects of activity are being tagged in a single log file entry. For example, in the E1 PTSBA certain symbols and spaces are recorded which are important for the underlying computational engine to understand the command but are not necessary for substantive analyses.

Process data such as log files thus typically require various steps of *preprocessing,* such as stemming and tagging, before they can be used for exploratory statistical

analyses. In *stemming*, all forms of the same command are replaced with the full command name. In *tagging*, each command is given a new label (i.e. *tag*) that reflects the purpose of the command; this can be seen as part of the process of *evidence identification*. Both procedures have their roots in the *natural language processing* literature [e.g. Feldman and Sanger 2006; Manning and Schuetze 1999].

Tagging log files carries with it a few important requirements that make this task harder in practice than it may seem intuitively. This is especially true for learning environments like *Packet Tracer* where tasks can be relatively unstructured with few natural boundaries unlike, say, different levels in a video game [e.g. Kerr et al. this issue]. Importantly, the tags for log files need to be interpretable from a substantive perspective by themselves, which can often be checked by asking subject-matter experts to 'read' tagged log file entries in order to see whether the resulting activity flow is interpretable [e.g. Gobert et al. this issue]. This ensures that interpretable descriptive analyses of log files can be done. Similarly, it ensures that any subsequent aggregations of log file entries or characterization of learner groups using summary statistics are also interpretable in practically meaningful terms.

Therefore, the labels can be neither too technical nor too detailed nor too simplistic – in other words, they needed to be at just the 'right' level of granularity for meaningful interpretability. Moreover, the process needs to be automated since a data set with complete log files of individual learners typically consists of thousands of entries. For example, for the E1 PTSBA exam, experts agreed that a minimum of 36 commands was required to successfully complete the task but learners produced up to 336 commands.

Through multiple conversations between subject-matter experts and statisticians we performed cycles of data analyses using various coding schemes and levels of granularity, some of which were retained and some of which were not. For example, we attempted to link log file entries to primary observables and compound observables directly. However, that proved difficult given that some learners showed rather inefficient solution behavior where it was difficult to associate a single log file entry with a particular action. In addition, some scores were given for tasks performed partially outside the log file environment (e.g. computing correct IP addresses) and so were not tracked directly in the log files. We also attempted to integrate a coding scheme associated with interface access levels on devices for some subsets of tasks but it proved

to be a far too technical to be of use for creating characterizations that could be easily summarized for learners and other users.

We eventually arrived at two tagging schemes that captured the functional objectives that the log file entries represent. At a coarse-grained level we differentiated between four types of log file entries, (a) management entries, (b) configuration entries, (c) verification entries, and (d) syntactically incorrect / 'fail' entries. At a fine-grained level we differentiated between 23 types of log file entries, which were related most closely to targeted actions of learners. These labels were closely related to the functional labels for the compound observables and performance components even though they are not fully identical with them due to the coding problems mentioned above; Figure 4 shows a part of a log file for a particular learner along with the coarse-grained and the fine-grained tags assigned to them.

	UserID	TimeStamp	IOS_Command_Device	IOS_Student_Command	Coarse_Grain_Tag	Fine_Grain_Tag
1	6535591	10:20:26	Router1	en	Management	Access Router
2	6535591	10:20:39	Router1	config t	Management	Change Command Levels
3	6535591	10:20:46	Router1	hostname Router1	Configure	Set Router Name
4	6535591	10:21:08	Router1	enable secret class	Configure	Set Router Password
5	6535591	10:21:14	Router1	lin con 0	Configure	Enter Console Config
6	6535591	10:21:24	Router1	login	Configure	Set Router Console Password
7	6535591	10:21:31	Router1	password cisco	Configure	Set Router Console Password
8	6535591	10:21:34	Router1	login	Configure	Set Router Console Password
9	6535591	10:21:46	Router1	lin vty 0 4	Configure	Enter Line Config
10	6535591	10:26:21	Switch1	password cisco	Configure	Set Switch Console Password
11	6535591	10:26:23	Switch1	login	Configure	Set Switch Console Password
12	6535591	10:26:31	Switch1	lin vty 0 15	Configure	Enter Line Config
13	6535591	10:26:37	Switch1	password cisco	Configure	Set Switch Line Password
14	6535591	10:26:40	Switch1	login	Configure	Set Switch Line Password
15	6535591	10:26:45	Switch1	exit	Management	Change Command Levels
16	6535591	10:29:53	PC1	ping 172.16.1.1	Verify	Verify host connectivity
17	6535591	10:43:08	PC1	ping 172.16.1.46	Verify	Verify host connectivity
18	6535591	10:47:24	PC2	ping 172.16.1.33	Verify	Verify host connectivity

Fig. 4. Sample log file entries with associated tags for a particular learner.

In the following section we describe how we used the coarse- and fine-grained tags as input for statistical analyses aimed at characterizing them in terms of the efficiency and effectiveness of their solution process, as mentioned at the beginning of this section.

4.2 Evidence Rules for Log Files

In many product data analyses the values of the observable variables are often binary reflecting a correct or an incorrect solution. However, there are no restrictions on the potential values that observable score variables for process data can assume; they simply must be of a format that can be included in a measurement model. On the basis of (a) the

time stamps for the log files as well as (b) the four coarse-grained tags and (c) the 23 fine-grained tags, we computed summary statistics for each learner's log file. The statistics included marginal measures such as the total time for the task and the total number of commands for the task as well as conditional tag-specific measures such as the total time, number of commands, and percentage of commands used per tag; we also included the total score on the E1 PTSBA in this set.

We then computed the distributions of these statistics across learners from the summative and formative subsamples; the distributional characteristics for key indicators and the coarse-grained tags only are shown in Table A-III. Clearly, the distributional features are similar across the two subsamples with the strongest notable differences being the larger maximum values and measures of variation for the formative subsample. However, these univariate distributions do not provide a direct insight into learner differences across subsamples.

Ideally, however, process data analysis should provide an added statistical value and added substantive insight beyond what one can learn from product data analysis. One way to investigate whether this is indeed the case is to examine the process data for individual learners who have identical or similar product scores. To illustrate this point, Table III displays select process summary scores for two learners from the summative subsample who both earned a total product score of 94 on the E1 PTSBA. Learner 1 used 49 commands to complete the task, had proportionally fewer device switches and management commands but more configuration commands than Learner 2; both had approximately the same proportion of fail and verification commands in alignment with their identical total scores.

Table III. Summary Statistics from Log Files for Two Learners with Identical Product Scores

	# of Commands	Time (*s* per command)	# of Device Switches	Configure (%)	Fail (%)	Manage (%)	Verify (%)
Learner 1	49	2357 (2s)	2	30 (61.2)	3 (6.0)	8 (16.3)	8 (16.3)
Learner 2	104	1964 (5s)	10	35 (33.6)	7 (6.7)	44 (42.3)	18 (17.3)

Note. *s* = seconds.

Clearly, these learners approached the task solution very differently even though they would both be characterized as equivalently competent if only their final submitted network configuration were used as evidence. Since the observable score variables for the log files can clearly provide some additional insight into learner performance, we sought to aggregate and synthesize the evidence contained within them next.

4.3 Measurement Model for Log Files

In order to combine the various summary statistics produced from the application of evidence rules to the log files, we performed PCAs with varimax rotations using the total score, the total number of commands, the total time on task, the number of switches between devices, and the percentage of commands for each of the four coarse-grained tags as outcome variables; we conducted two separate PCAs for the two learner subsamples. For both subsets, we were able to reliably extract two components that accounted for about 60% of the total variance in the log file variables; we subsequently saved the estimated PCA scores on both components using the regression method.

Table IV shows the varimax rotated loading matrices for the formative and summative learner subsamples, which show very similar loading patterns across the two learner subsamples; recall that the loadings for the individual variables are their correlations with the dimensional subscores created by the PCA.

Table IV. Loadings for the First Two Principal Components for Subsamples

	Formative Subsample		Summative Subsample	
	Component 1 (*Inefficiency*)	Component 2 (*Effectiveness*)	Component 1 (*Inefficiency*)	Component 2 (*Effectiveness*)
Total score	.300	**.756**	-.180	**.850**
# of commands	**.864**	-.103	**.828**	.260
Total time	**.609**	**-.381**	**.677**	-.168
# of switches	**.839**	.104	**.757**	**.382**
% manage	**.397**	.166	.268	**.531**
% configure	-.340	**.780**	**-.781**	.178
% verify	**.427**	.309	.205	**.387**
% fail	-.140	**-.929**	**.408**	**-.843**

Notes. Loadings greater than .35 are shown in boldface. Time is measured in seconds; percentages are computed based on all commands for each learner.

For both learner subsamples, score variation on the first principal component was consistently strongly driven by the total number of commands, the total time on task, and the number of switches between devices; these loadings are indicative of a measure of *inefficiency*. To make this dimension more easily interpretable vis-à-vis the effectiveness dimension, we reversed the sign on the saved scores so that higher scores reflected, indeed, a higher degree of *efficiency* in the task solution.

Similarly, for both learner subsamples, score variation on the second principal component was consistently strongly driven by the total score and the percentage of fail commands; these loadings are indicative of a measure of *effectiveness*. The signs of the loadings on the effectiveness component are in line with what would be expected in that learners with higher total scores and a lower percentage of fail commands have a higher score on this component.

The percentage of verify commands has relatively moderate loadings on both components across both subsamples, suggesting that it is not as informative in distinguishing between efficient or effective learners. For both subsamples, the percentage of management and configuration commands is slightly more informative about learner differences. For example, for the formative subset the percentage of configuration commands loads strongly positively on the effectiveness component while for the summative subset it loads strongly negatively on the inefficiency component. Similarly, the percentage of management commands loads moderately positively on the inefficiency component for the formative subsample but moderately positively on the effectiveness component for the summative subsample.

This seemingly divergent pattern is congruent with the expected behavior within each subsample upon reflection, however as there is a wider range of ability in the formative subsample. Therefore, for the formative subsample higher percentages of configuration and management commands indicate a more effective, albeit somewhat less efficient, solution relative to other learners in that subsample. In the summative subsample, higher percentages of configuration and management commands indicate a more effective and efficient solution relative to other learners in that subsample, indicating that these actions are likely more targeted in this subsample.

In order to connect the two-dimensional scores of the learners with the proficiency / primary effectiveness characterizations from the product data, we used the latent class assignments from the diagnostic measurement analyses for the component-based scoring

structure. As mentioned at the end of the second main section in this paper, we chose the ordered five-class solution as a conditioning variable. We focused on the DCM analyses because (a) they were successfully completed for both the summative and formative subsets and (b) they provided a very similar latent class assignment as the Bayes net analysis for the summative subset.

Recall that the relationship between the latent-class membership variable and the total score on the assessment was rather strong as illustrated in Figure 3 earlier in the paper along with the associated effect-size measures. Thus, conditioning on the latent class membership for this particular DCM model structure can be methodologically seen as a statistically 'refined' version of conditioning on proficiency score bands.

Figure A5 shows the distributions of the effectiveness and efficiency scores from the PCAs separately across the two learner subsamples. We can see that the locations of the score distributions for the effectiveness dimension increase with latent classes that represent the mastery of more skills across both learner subsamples, as one would expect. A one-way ANOVA for the effectiveness score using latent class membership as the independent variable and post-hoc polynomial contrasts confirmed that the mean effectiveness score followed a monotonically increasing cubic trend across the latent classes for the formative population (t = .210, p < .001, partial η^2 = .473) as well as the summative population (t = .369, p < .001, partial η^2 = .584).

In contrast, the locations of the score distributions for the efficiency dimension decrease across latent classes. A one-way ANOVA for the efficiency score using latent class membership as the independent variable and post-hoc polynomial contrasts confirmed that the mean efficiency score followed a weak quadratic trend across the latent classes for the formative population (t = .249, p < .001, partial η^2 = .107) and showed no specific reliable trend for the summative population (F(4,272) = 1.593, n.s., partial η^2 = .023).

In addition, the range of the efficiency scores increased across the five latent classes for both learner subsamples while the range of the effectiveness scores decreased. A plausible interpretation for these patterns is as follows. Even though the efficiency score distribution for the first latent class was located higher than those of the other latent classes and showed less variance, we know that this is a class of weak learners, which implies that the efficiency score should be more appropriately interpreted as a measure of 'brevity' for these learners. In other words, higher efficiency scores have a negative

practical connotation for weaker learners since they have lower proficiency scores overall, but they have a positive practical connotation for stronger learners who have higher proficiency scores overall.

A further inspection of the bivariate distributions for the effectiveness and efficiency scores across the latent classes revealed an even more differentiated picture; Figure A6 shows these relationships for the larger formative subset. For the first two latent classes, which represent those learners who have either not mastered any skills or who have mastered only DC, the two scores remain essentially uncorrelated. For the other three latent classes, which represent learners with higher degrees of mastery, there is a positive relationship, albeit only a moderately strong one, between effectiveness and efficiency.

That is, for each of these three different subgroups of learners, the more effective learners are also the more efficient ones, on average. Moreover, as noted earlier, learners are placed in higher score ranges on the effectiveness dimension as reflected by the relative positioning of the bivariate scatterplots. The strength of this bivariate relationship is relatively weak, however, with R^2 values ranging from about 15 to 23 percent. It will be important to investigate such patterns with future samples of learners to see whether stronger relationships become visible.

Finally, returning to our two learners from Table III, we examined the efficiency and effectiveness scores for both. As one might expect, the efficiency score of Learner 1 is .75 while the score for Learner 2 is -.99, indicating that Learner 1 is much more efficient in his or her solution process than Learner 2 while both learners have a similar effectiveness score.

In summary, in the previous two subsections we provided an illustration of how summary statistics can be used to create observable variables from log files. We then demonstrated how an evidence accumulation or measurement model was developed to synthesize these observable variables into measures of effectiveness and efficiency. These measures revealed differences in the performance of learners whose proficiency scores from the product data indicated that they performed similarly. It revealed that learners can vary in the efficiency with which they execute the task, even when their total score or latent class membership is similar or even identical. More importantly, it revealed that the relation between efficiency and effectiveness in task solution varies in ways that can be meaningfully interpreted, which conceptually opens up avenues for designing interventions targeted to specific subpopulations of learners.

In closing, we note one important caveat that we alluded to at the beginning of the paper, which is the partial representativeness of our sample of learners. Thus, if the heterogeneity in the population influences the response processes to the E1 PTSBA differentially (i.e. if different DCM or Bayes net models or PCA solutions would be appropriate for different subpopulations defined by geographical region) then the results in this paper would not be generalizable to the overall population of learners in the *Networking Academy*. More data are needed to investigate this question in more depth but some caution with regards to the generalization of our findings is probably warranted.

5. DISCUSSION

In this paper we have described various key steps in the creation of evidence models using evidence rules and statistical measurement models for product and process data that result from the *Packet Tracer* digital learning environment in the Cisco *Networking Academy*. We want to close this paper with a few reflections on the lessons that we have learned in the process that we think are useful for others to consider as well.

5.1 On Evidence Models for Product Data

Though much of statistical modeling of assessment data in educational measurement is geared toward the identification of a single 'correct' model, our illustrations in this paper showed that this need not be the default. Multiple models may not only be used to gain an understanding of different facets of the data but may also be used to support different reporting levels for different use contexts. In ECD parlance, this means that it can be perfectly defensible to have several different student models tied to different evidence models.

For example, if we desire to characterize learners in terms of the four particular subskills that motivated, in part, the components-based scoring structure this would lead us to use statistical models that reflect that structure as we did in our DCM and Bayes net analyses. A desire to explicitly link performance on the *Packet Tracer* assessment with other assessments developed using theoretical claims about learner competencies would lead to the use of the claims-based scoring model and associated statistical models. The desire for a single proficiency score, rather than a multidimensional score profile, might lead us to use a different aggregation (e.g. the unidimensional proficiency score from IRT or CTT or a coarse-grained proficiency classification in a higher-order DCM or a four-level Bayes net).

This openness to considering and using multiple measurement models does not mean that any statistical model is empirically defensible, of course. An honest analytical critique, like the ones that we conducted in the second main section of this paper, highlights its strengths in terms of reproducing certain aspects of the observed data patterns in a consistent manner and, more importantly perhaps, its weaknesses. This can lead to valuable insights that would be lost if one only searched for a single true model while discarding any strongly misfitting models during this process. As we noted in the second section of the paper as well, criticism of the model-data fit for various models associated with the component-based scoring structure for the product data motivated the creation of a second claims-based scoring structure, which possessed procedural validity and resulted in superior model-data fit in various instances.

Philosophically speaking, this line of argumentation reflects a distinct frame of thinking about what modeling is supposed to accomplish, because it reflects a belief that the development of an assessment and associated statistical model(s) is not about seeking the single true structure of the world. As Box [1976] noted, this is a fool's errand, for all our models are wrong in the end. Rather, a model is built by an interdisciplinary team of researchers to capture certain relevant features of the data while necessarily ignoring others in the service of desired inferences.

5.2 On Leveraging Collateral Information about Learners

As noted throughout this paper, analyses of product and process data from digital learning environments for the purpose of diagnostic assessment are complicated by the openness of the workspace and the resulting variability in learner behaviors. In order to build a comprehensive validation argument for inferences about learners [Kane 2006], analysts must leverage a wide variety of data sources based on results from domain modeling and analysis steps in the ECD framework, including collateral information beyond the product and process data learners produce.

In the case of the E1 PTSBA, sampling considerations suggested the distinction between the summative and formative subsample, which led to differential hypotheses about performance that were largely corroborated using the product data. Similarly, additional collateral information could take the form of scores from other assessments that learners have been taking. In the case of the *Networking Academy*, performance information from traditional standardized assessments could be used as covariates in

explanatory modeling approaches or to inform the structure of particular models that are being used (e.g. disallowing particular latent classes in DCMs or Bayes nets). Thus, it is not just combining information from multiple assessments that matters, but also leveraging information from one established assessment to help understand modeling approaches for novel assessments such as the E1 PTSBA. While this work was beyond the scope of this paper, it is part of our research agenda for the project.

Incorporating collateral information about learners can also be viewed as an instance of expanding the conceptual narrative of the model [Mislevy et al. 2008]. As illustrated in this paper, a straightforward analytic approach to multiple subgroups involves conducting separate analyses for each subgroup. Additional model-based approaches include multi-group models where group membership is known and latent mixture models where group membership is unknown. Such approaches allow for the simultaneous analysis of all data and provide procedures for addressing questions such as parameter invariance and group differences in latent variables. Extensions of these models exist that allow for the incorporation of collateral information about tasks and learners, an area known as *explanatory (item response) modeling* [e.g. De Boeck and Wilson 2004; Mislevy et al. 2008].

Similarly, new models might be created by combining various features of certain existing models for suiting the comprehensive inferential purpose of a program such as the *Networking Academy*. This requires a flexible and powerful approach to modeling of which the Bayesian modeling paradigm is an attractive example. In this paper, we discussed the use of a fully Bayesian approach for fitting Bayes nets to both subsamples; the framework can also be used to incorporate many additional complex features including multilevel structures, latent mixtures, collateral covariates, and missing data [Levy 2009; Levy et al. 2011; Lynch 2007]. In addition, a Bayesian approach is aligned with the principles of ECD in terms of the nature and form of the assessment argument and the use of probability-based reasoning [Mislevy and Levy 2007; Mislevy et al. 2003].

5.3 On Evidence Models for Process Data

Due to the volume of data produced, we saw that making use of process data required automation of procedures. With digital environments such as *Packet Tracer*, the potential to record activity sequences is built in, but the critical issue becomes knowing what to

record and how to make sense of it. Here is where design-based thinking, guided by a principled development framework such as ECD, becomes important. Decisions about what features of behavior to pay attention to, and how they have evidentiary bearing on the desired inferences, require an integration of expert knowledge from areas as diverse as the targeted professional domain, curriculum development, instructional design, the creation of digital learning tools, the learning sciences more broadly, computer programming, multivariate statistics, and modern measurement.

In our project, the specification and implementation of evidence rules for process data consumed far more time than the specification and implementation of evidence rules for product data. Much of the discussion of the analyses of the process data involved the tagging of log file entries. As noted in the previous section, tagging is a form of evidence identification where we take something the learner does and identify the relevant aspects of performance that we want to pay attention to. In our case, it was the function of the commands as reflected by the two sets of tags at different levels of grain size.

No matter what aspect of the data structure we are shining an evidentiary spotlight on, however, there are always also aspects that we are leaving in the dark. To the extent that we are blending out things that are not of inferential interest to us, this is not a problem. To the extent that what we neglect is indeed relevant to our desired inferences, this becomes a problem. To return to the case study of our learners in the previous section who scored 94 on the E1 PTSBA, if we only cared to make inferences or distinctions among students based on the functionality of the resulting network (i.e. the effectiveness of their solution), the process data are not needed at all. In fact, all distinctions are perfectly captured by the total score on the assessment and the dimensional subscores. However, if we are interested in notions of efficiency as well, then it is important to recognize that learners who have the same scores in terms of the product data might vary wildly in terms of their efficiency. We conjecture that most researchers are interested in some version of a two-axial characterization of learners based on efficiency and effectiveness.

5.4 On the Linearity of ECD Representations vs. Nonlinearity of Its Implementation

The power of the ECD framework lies in its ability to frame the discussion about the desired inferences / distinctions we want to make about learners and how to make sense

of their complex behavior to make those inferences / distinctions. Even though there may seem to be a linear flow suggested by many graphical representations of the ECD framework, as we have illustrated in this paper, in practice the constituent processes are decidedly nonlinear [Mislevy et al. 2011, this issue]. The process is not merely nonlinear at any given point in time for a given purpose, but the way we think about evidence identification and accumulation within an ECD framework also changes when data-analysis purposes are changed.

For example, the operational use of a scoring structure based on product data is aligned with a set of desired inferences and distinctions about learners based on the accuracy of the final network configuration. An additional desire to pursue learner distinctions around a construct like efficiency brings on new questions about acceptable behavior in the domain and possible sources of evidence for such behavior. For example, do we operationalize efficiency simply by the time taken for the task, more complex aggregates based on multiple variables, more linear solution processes, or solution processes that are more closely aligned with particular prototypical ones taught in the classroom? Novel evidentiary desiderata often bring with them consequences for the design of new assessments and the re-design of existing assessments.

5.5 On the Interplay between Subject-matter Experts and Statisticians

As we have clearly shown repeatedly in this paper, having either just data or just theory is not enough to create meaningful inferences. For example, scoring structures for final network configurations and tags for log files do not merely 'emerge' from the data or are a priori 'conceived of'; instead, they are equally informed by expert consensus and by results from exploratory and confirmatory statistical analyses. Put differently, it is the interplay between theory and data that is most powerful. As we have demonstrated in this paper, just the specification of the ECD evidence model requires both a large tool kit of potential analysis methods and collaboration among experts in a variety of fields.

While this interplay between theory and data holds much potential for improving diagnostic assessment within digital learning environments, it is not always simple to implement in practice and we have encountered logistical challenges in this project that are probably familiar to researchers working in similar contexts. To name a few of these, we have observed that different resource constraints across different teams such as limited financial resources and time can make continual information exchanges difficult at times, operational requirements for updating a computational engine can make

implementing suggested changes infeasible or impractical, the scope of use of a learning environment can make the tracking of rich collateral information prohibitive and impractical, and a lack of well-understood guidelines for desired levels of reporting for different subpopulations of learners can make it difficult to justify certain operational changes to the computational engine. Finally, changes in evidence identification and accumulation procedures may also hold implications for the design of future assessments, and thus have implications for the development of conceptual models of learning over time that can be cost-intensive to implement.

Put differently, interdisciplinary research teams that work with digital learning environments for diagnostic assessment purposes have to decide what an appropriate balance is between efforts that lead to changes in operational practice and efforts that lead to the creation and documentation of a comprehensive program of validation [e.g. Kane 2006; Williamson et al. 2012]. Even though they are both obviously interconnected, the associated allocation of resources will likely differ notably depending on which goal is pursued at any given point in time. As we have demonstrated in this paper, the ECD framework can guide such decision-making processes. Furthermore, the specification, implementation, and empirical evaluation of evidence rules and measurement models for product and process data play an important empirical part in this complex endeavor.

ACKNOWLEDGEMENTS

This research was supported by Cisco. The work of Dr. Mislevy 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 work of Dr. Rupp was supported, in part, by two grants from the National Science Foundation awarded to the University of Wisconsin at Madison (DRL-0918409 and DRL-0946372). 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), the National Science Foundation (NSF), or the U.S. Department of Education.

REFERENCES

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 1*(5). Available online at http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml

ALMOND, R. G., WILLIAMSON, D. M., MISLEVY, R. J., AND YAN, D. in press. *Bayes nets in educational assessment*. Springer, New York, NY.

BAKER, R. S. J. D., AND YACEF, K. 2009. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3-17.

BOX, G. E. 1976. Science and statistics. *Journal of the American Statistical Association 71,* 791-799.

CHAPPLE, K., JOHNSON, A., ROSSON, J., WEST, P., STANLEY, K., TERMAAT, B., AND BEHRENS, J. T. 2009. Applying cross-disciplinary threads to a global assessment system with an emphasis on simulation and gaming. Paper presented at the *Annual meeting of the American Educational Research Association (AERA)*, San Diego, CA.

CROCKER, L., AND ALGINA, J. 1986. *Introduction to classical and modern test theory*. Wadsworth, Belmont, CA.

DE BOECK, P., AND WILSON, M., Eds. 2004. *Explanatory item response models: A generalized linear and nonlinear approach.* Springer, New York, NY.

FELDMAN, R., AND SANGER, J. 2006. *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press, Cambridge, UK.

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, DC, 555– 560.

FREZZO, D. C., BEHRENS, J. T., MISLEVY, R. J. 2010. Design patterns for learning and assessment: Facilitating the introduction of a complex simulation-based learning environment into a community of instructors. *Journal of Science Education and Technology 19*, 105-114.

GOBERT, J. D., SAO PEDRO, M. A., BAKER, R. S. J. D., TOTO, E., AND MONTALVO, O. this issue. Leveraging educational data mining for real time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining*.

JOLIFFE, I. T. 2010. *Principal component analysis, 2nd ed*. Springer, New York, NY.

KANE, M. 2006. Validation. In *Educational measurement,* 4th ed., R. L. BRENNAN, Ed. American Council on Education / Praeger, Washington, DC, 18-64.

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

KLINE, T. 2005. *Psychological testing: A practical approach to design and evaluation*. Sage, Thousand Oaks, CA.

KUNINA-HABENICHT, O., RUPP, A. A., AND WILHELM, O. 2012. The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement 49*, 59-81.

LATTIN, J., CARROLL, D., AND GREEN, P. 2002. *Analyzing multivariate data*. Duxbury Press, New York, NY.

LEIGHTON, J. P. 2004. Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice 23*(4), 6-15.

LEVY, R. 2009. The rise of Markov chain Monte Carlo estimation for psychometric modeling. *Journal of Probability and Statistics*, Article ID 537139, 18 pages.

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

LEVY, R., CRAWFORD, A. V., FAY, D., AND POOLE, K. L. 2011. Data-model fit assessment for Bayesian networks for simulation-based assessments. Presented at the *Annual meeting of the American Educational Research Association (AERA)*, New Orleans, LA.

LEVY, R., MISLEVY, R. J., AND BEHRENS, J. T. 2011. MCMC in educational research. In *Handbook of Markov chain Monte Carlo: Methods and applications,* S. BROOKS, A. GELMAN, G. L. JONES, AND X. L. MENG, Eds., Chapman and Hall/CRC, London, UK, 531-545.

LEVY, R., AND SVETINA, D. 2011. A generalized dimensionality discrepancy measure for dimensionality assessment in multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology 64,* 208-232.

LYNCH, S. 2007. *Introduction to applied Bayesian statistics and estimation for social scientists.* Springer, New York, NY.

MANNING, C. D., AND Schuetze, H. 1999. *Foundations of statistical natural language processing*. MIT Press, Boston, MA.

MARIS, E. 1999. Estimating multiple classification latent class models. *Psychometrika 64*, 187-212.
MISLEVY, R. J., BEHRENS, J. T., LEVY, R., AND DICERBO, K. E. 2011. *The interplay of design and data exploration in an evolving assessment system*. Manuscript submitted for publication.
MISLEVY, R. J., AND LEVY, R. 2007. Bayesian psychometric modeling from an evidence-centered design perspective. In *Handbook of statistics, volume 26: Psychometrics,* C. R. RAO AND S. SINHARAY, Eds., North-Holland, Amsterdam, The Netherlands, 839-865.
MISLEVY, R. J., LEVY, R., KROOPNICK, M., AND RUTSTEIN, D. 2008. Evidentiary foundations of mixture item response theory models. In *Advances in latent variable mixture models,* G. R. HANCOCK AND K. M. SAMUELSEN, Eds., Information Age Publishing, Charlotte, NC, 149-175.
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.
MISLEVY, R. J., STEINBERG, L. S., ALMOND, R. G., AND LUKAS, J. F. 2006. Concepts, terminology, and basic models of evidence-centered design. In *Automated scoring of complex tasks in computer-based testing*, D. M. WILLIAMSON, R. J. MISLEVY, AND I. I. BEJAR, Eds. Erlbaum, Mahway, NJ.
MISLEVY, R. J., BEHRENS, J. T., DICERBO, K. E., AND LEVY, R. this issue. Design and discovery in educational assessment: Evidence centered design, psychometrics, and data mining. *Journal of Educational Data Mining*.
RECKASE, M. D. 2009. *Multidimensional item response theory*. Springer, New York, NY.
RAYKOV, T., AND MARCOULIDES, G. A. 2011. *Introduction to psychometric theory*. Taylor and Francis, New York, NY.
ROMERO, C., VENTURA, S., PECHENIZKIY, M., AND BAKER, R. S. J. D., Eds. 2010. *Handbook of educational data mining*. Chapman and Hall / CRC, New York, NY.
RUPP, A., AND TEMPLIN, J. 2008. Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*: *Interdisciplinary Research and Perspectives 6*, 219-262.
RUPP, A. A., TEMPLIN, J., AND HENSON, R. A. 2010. *Diagnostic measurement: Theory, methods, and applications*. Guilford Press, New York.
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://ejournals.bc.edu/ojs/index.php/jtla/article/view/1623
SCOTT, J. P., AND CARRINGTON, P. 2011. *The SAGE handbook of social network analysis*. Sage, Thousand Oaks, CA.
SINHARAY, S. 2006. Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics 31,* 1-33.
TATSUOKA, K. K. 2009. *Cognitive assessment: An introduction to the rule-space method*. Routledge, Florence, KY.
THISSEN, D., AND WAINER, H. Eds. 2001. *Test scoring*. Mahwah, NJ: Erlbaum.
VAN DER AALST, W. M. P. 2011. *Process mining: Discovery, conformance and enhancement of business processes*. Springer, New York, NY.
WILLIAMSON, D. M., XI, X., AND BREYER, F. J. 2012. A framework for the evaluation and use of automated scoring. *Educational Measurement: Issues and Practice 31(1)*, 2-13.
YEN, W. M. 1984. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement 8*, 125-145.

Received August 2011; revised July 2012; accepted July 2012.

APPENDIX

Table A-I. Descriptive Statistics for Primary Observables from the Final Network Configuration

PC	ID	Device	Command	Classical Test Theory Statistics									
				Difficulty		Discrimination (Total Score)				Discrimination (Subscores)			
				S	F	S	F	S	F	S	F	S	F
				(p)	(p)	(pbs)	(pbs)	(D)	(D)	(pbs)	(pbs)	(D)	(D)
DC	1	PC 1	Link to Switch 1	.91	.92	.41	.41	.22	.22	.62	.62	.33	.29
DC	2	Router 1	Link to Switch 1	.94	.92	.56	.44	.19	.26	.99	.85	.24	.31
DC	3	Router 1	Link to Switch 2	.95	.93	.57	.49	.18	.28	.77	.85	.20	.27
IP	4	Router 1	Power (0/0)	.96	.89	.64	.79	.12	.65	.62	.79	.12	.37
IP	5	Router 1	IP Address (0/0)	.91	.81	.70	.76	.29	.83	.72	.82	.28	.62
IP	6	Router 1	Subnet Mask (0/0)	.91	.79	.78	.82	.32	.89	.79	.87	.29	.70
IP	7	Router 1	Subnet Mask (0/1)	.96	.84	.65	.81	.14	.80	.65	.83	.14	.52
IP	8	Router 1	Power (0/1)	.86	.70	.76	.79	.44	.95	.78	.84	.40	.85
IP	9	Router 1	IP Address (0/1)	.88	.72	.73	.78	.39	.93	.69	.80	.33	.76
IP	10	Router 1	Description (0/0)	.79	.59	.77	.62	.64	.80	.74	.65	.70	.75
IP	11	Router 1	Description (0/1)	.78	.56	.75	.66	.65	.81	.73	.69	.73	.79
IP	12	Switch 1	Power	.88	.69	.61	.73	.32	.88	.62	.74	.32	.72
IP	13	Switch 1	IP Address	.88	.69	.85	.81	.44	.95	.86	.86	.41	.85
IP	14	Switch 1	Subnet Mask	.87	.69	.80	.83	.44	.96	.87	.87	.41	.88
IP	15	Switch 1	Default Gateway	.70	.48	.69	.74	.70	.80	.76	.78	.57	.75
IP	16	PC 1	IP Address	.86	.71	.82	.73	.51	.82	.80	.79	.43	.79
IP	17	PC 1	Subnet Mask	.88	.74	.70	.73	.40	.83	.74	.77	.36	.71
IP	18	PC 1	Default Gateway	.85	.74	.72	.69	.43	.75	.72	.73	.41	.68

BDC	19	Router 1	Host Name	.94	.88	.60	.64	.19	.59	.65	.68	.15	.32
BDC	20	Router 1	Enable Secret	.85	.71	.55	.65	.38	.80	.64	.73	.26	.66
BDC	21	Router 1	Login (console)	.78	.63	.55	.61	.52	.80	.77	.75	.59	.74
BDC	22	Router 1	Password (console)	.71	.61	.61	.68	.64	.89	.81	.82	.67	.81
BDC	23	Router 1	Login (VTY)	.78	.64	.62	.63	.61	.81	.76	.75	.56	.74
BDC	24	Router 1	Password (VTY)	.78	.63	.61	.67	.57	.87	.72	.79	.53	.79
BDC	25	Router 1	Banner motd	.93	.79	.65	.69	.23	.79	.69	.74	.17	.54
BDC	26	Switch 1	Host Name	.92	.80	.59	.75	.23	.86	.64	.78	.17	.57
BDC	27	Switch 1	Enable Secret	.83	.67	.62	.74	.45	.92	.68	.79	.28	.81
BDC	28	Switch 1	Login (Console)	.73	.56	.69	.75	.64	.94	.84	.86	.66	.92
BDC	29	Switch 1	Password (Console)	.69	.56	.74	.77	.71	.97	.87	.88	.71	.93
BDC	30	Switch 1	Login (VTY)	.75	.57	.71	.76	.68	.95	.82	.87	.61	.92
BDC	31	Switch 1	Password (VTY)	.76	.60	.68	.77	.65	.97	.74	.86	.56	.90
BDC	32	Switch 1	Banner motd	.91	.72	.65	.74	.27	.91	.66	.77	.19	.71
VT	33	PC 2	IP Address	.76	.56	.78	.76	.69	.88	.90	.93	.62	.99
VT	34	PC 2	Subnet Mask	.79	.62	.68	.75	.57	.88	.84	.89	.53	.96
VT	35	PC 2	Default Gateway	.70	.54	.74	.76	.71	.88	.87	.88	.77	.97
VT	36	PC 2	Link to Switch 2	.89	.77	.53	.58	.31	.60	.68	.80	.29	.58

Note. PC = performance component, DC = device connection, IP = IP addressing, BDC = basic device configuration, VT = verification and troubleshooting, F = formative subsample, S = summative subsample, pbs = point-biserial correlation coefficient, p = p-value / % correct, D = discrimination coefficient (i.e., differences in p-values between low- and high-scoring groups).

Table A-II. Postulated Loading Structure and CTT Discriminations for Claims-based Scoring Structure

Claim(s)	PO	Device	Command	Claim 3			Claim 4				Claim 5				Claim 3		Claim 4		Claim 5	
				3.2	3.3	3.5	4.2	4.4	4.5	4.6	5.1	5.2	5.3	5.4	S	F	S	F	S	F
5	1	PC 1	Link to Switch 1	0	0	0	0	0	0	0	0	0	1	0	.28	.20	.28	.23	.33	.27
5	2	Router 1	Link to Switch 1	0	0	0	0	0	0	0	0	0	1	0	**.44**	**.23**	**.43**	**.24**	**.49**	**.30**
5	3	Router 1	Link to Switch 2	0	0	0	0	0	0	0	0	0	1	0	**.45**	**.24**	**.42**	**.26**	**.49**	**.32**
4, 5	4	Router 1	Power (0/0)	0	0	0	1	0	1	0	1	1	0	0	.48	.52	.56	.61	.55	.61
3, 4, 5	5	Router 1	IP Address (0/0)	1	1	0	1	0	1	0	1	1	0	0	.74	.74	.62	.64	.66	.67
3, 4, 5	6	Router 1	Subnet Mask (0/0)	1	1	0	1	0	1	0	1	1	0	0	.79	.80	.69	.70	.72	.73
3, 4, 5	7	Router 1	Subnet Mask (0/1)	1	1	0	1	0	1	0	1	1	0	0	.74	.76	.65	.68	.69	.71
4, 5	8	Router 1	Power (0/1)	0	0	0	1	0	1	0	1	1	0	0	**.52**	**.63**	**.57**	**.67**	**.57**	**.69**
3, 4, 5	9	Router 1	IP Address (0/1)	1	1	0	1	0	1	0	1	1	0	0	.84	.83	.66	.68	.72	.73
4, 5	10	Router 1	Description (0/0)	0	0	0	1	0	1	0	1	1	0	0	**.53**	**.39**	**.72**	**.55**	**.70**	**.53**
4, 5	11	Router 1	Description (0/1)	0	0	0	1	0	1	0	1	1	0	0	.52	.42	**.70**	**.57**	.68	.56
4, 5	12	Switch 1	Power	0	0	0	1	0	1	0	1	1	0	0	**.50**	**.62**	**.56**	**.66**	**.57**	**.67**
3, 4, 5	13	Switch 1	IP Address	1	1	0	1	0	1	0	1	1	0	0	.81	.79	.73	.73	.77	.75
3, 4, 5	14	Switch 1	Subnet Mask	1	1	0	1	0	1	0	1	1	0	0	.82	.81	.71	.74	.75	.76
3, 4, 5	15	Switch 1	Default Gateway	1	1	0	1	0	1	0	1	1	0	0	.65	.63	.57	.59	.60	.60
3, 5	16	PC 1	IP Address	1	1	0	0	0	0	0	0	0	0	1	.86	.79	.67	.59	**.77**	**.67**
3, 5	17	PC 1	Subnet Mask	1	1	0	0	0	0	0	0	0	0	1	.77	.75	.61	.59	.68	.66
3, 5	18	PC 1	Default Gateway	1	1	0	0	0	0	0	0	0	0	1	.74	.70	.60	.54	.68	.61

4, 5	19	Router 1	Host Name	0	0	0	1	1	0	0	1	0	0	0	.41	.39	.53	.53	.52	.51
4, 5	20	Router 1	Enable Secret	0	0	0	1	0	0	1	1	0	0	0	.35	.42	.51	.61	**.41**	**.58**
4, 5	21	Router 1	Login (con)	0	0	0	1	0	0	1	1	0	0	0	.26	.31	.56	.57	.48	.53
4, 5	22	Router 1	Password (con)	0	0	0	1	0	0	1	1	0	0	0	**.28**	**.38**	.59	.63	.53	.59
4, 5	23	Router 1	Login (VTY)	0	0	0	1	0	0	1	1	0	0	0	.38	.35	.62	.58	.56	.55
4, 5	24	Router 1	Password (VTY)	0	0	0	1	0	0	1	1	0	0	0	.41	.40	.57	.63	.55	.59
4, 5	25	Router 1	Banner motd	0	0	0	1	0	0	1	1	0	0	0	.52	.47	.62	.65	.61	.62
4, 5	26	Switch 1	Host Name	0	0	0	1	1	0	0	1	0	0	0	**.43**	**.54**	**.57**	**.69**	**.55**	**.67**
4, 5	27	Switch 1	Enable Secret	0	0	0	1	0	0	1	1	0	0	0	**.40**	**.50**	**.58**	**.69**	**.56**	**.66**
4, 5	28	Switch 1	Login (con)	0	0	0	1	0	0	1	1	0	0	0	.36	.43	.66	.68	.60	.64
4, 5	29	Switch 1	Password (con)	0	0	0	1	0	0	1	1	0	0	0	**.37**	**.46**	.68	.70	.63	.66
4, 5	30	Switch 1	Login (VTY)	0	0	0	1	0	0	1	1	0	0	0	.44	.45	.68	.70	.63	.66
4, 5	31	Switch 1	Password (VTY)	0	0	0	1	0	0	1	1	0	0	0	.46	.48	.63	.71	.60	.67
4, 5	32	Switch 1	Banner motd	0	0	0	1	0	0	1	1	0	0	0	.51	.53	.61	.70	.60	.67
3, 5	33	PC 2	IP Address	1	1	1	0	0	0	0	0	0	0	1	.80	.77	.61	.56	.71	.66
3, 5	34	PC 2	Subnet Mask	1	1	1	0	0	0	0	0	0	0	1	.72	.74	.54	.58	.63	.66
3, 5	35	PC 2	Default Gate	1	1	1	0	0	0	0	0	0	0	1	.74	.75	.53	.55	.64	.64
5	36	PC 2	Link to Switch 2	0	0	0	0	0	0	0	0	0	1	0	.45	.48	.40	.42	.48	.50

Notes. PO = primary observable, F = formative subsample, S = summative subsample, con = console. Discrimination index values are the point-biserial correlation coefficients using the three separate dimensional total scores for claims dimension 3, 4, and 5, respectively; all correlations were significant at the $p < .01$ level. Differences between correlations for summative and formative subsample greater than or equal to .10 are shown in boldface.

Table A-III. Descriptive Statistics for Log Files for Learner Subsamples

	Summative Subsample (n = 277)						Formative Subsample (n = 1,942)					
	Min	Max	Mean	Median	IQR	SD	Min	Max	Mean	Median	IQR	SD
Total score	4.00	100.00	80.56	90.00	27.00	24.15	0.00	100.00	64.53	72.00	48.00	29.02
Total time (in seconds)	594.00	6659.00	2378.37	2155.00	1384.00	1103.93	311.00	7182.00	2593.59	2307	1869.00	1418.24
# of commands	12.00	235.00	90.14	81.00	48.00	39.06	2.00	336.00	80.77	72.00	47.00	41.56
Mean time per command	0.00	0.12	0.04	0.04	0.02	0.02	0.00	0.32	0.04	0.03	0.02	0.02
# of switches b/t devices	0.00	25.00	6.94	6.00	5.00	4.49	0.00	46.00	5.63	4.00	5.00	4.86
# of manage commands	0.00	97.00	29.18	24.00	21.00	17.57	0.00	122.00	24.57	21.00	20.00	17.28
# of configure commands	0.00	78.00	35.52	34.00	9.00	11.63	0.00	166.00	31.32	31.50	12.00	14.29
# of verify commands	0.00	50.00	11.74	10.00	10.00	8.85	0.00	92.00	8.86	7.00	9.00	9.23
# of fail commands	0.00	90.00	13.90	10.00	14.00	13.94	0.00	174.00	15.26	10.00	15.00	16.46
% of manage commands	0.00	0.59	0.31	0.31	0.12	0.09	0.00	0.77	0.30	0.29	0.12	0.09
% of configure commands	0.00	0.67	0.41	0.43	0.17	0.11	0.00	0.78	0.40	0.42	0.19	0.15
% of verify commands	0.00	0.60	0.13	0.12	0.10	0.08	0.00	1.00	0.10	0.09	0.11	0.09
% of fail commands	0.00	0.80	0.15	0.12	0.13	0.12	0.00	1.00	0.19	0.14	0.16	0.17

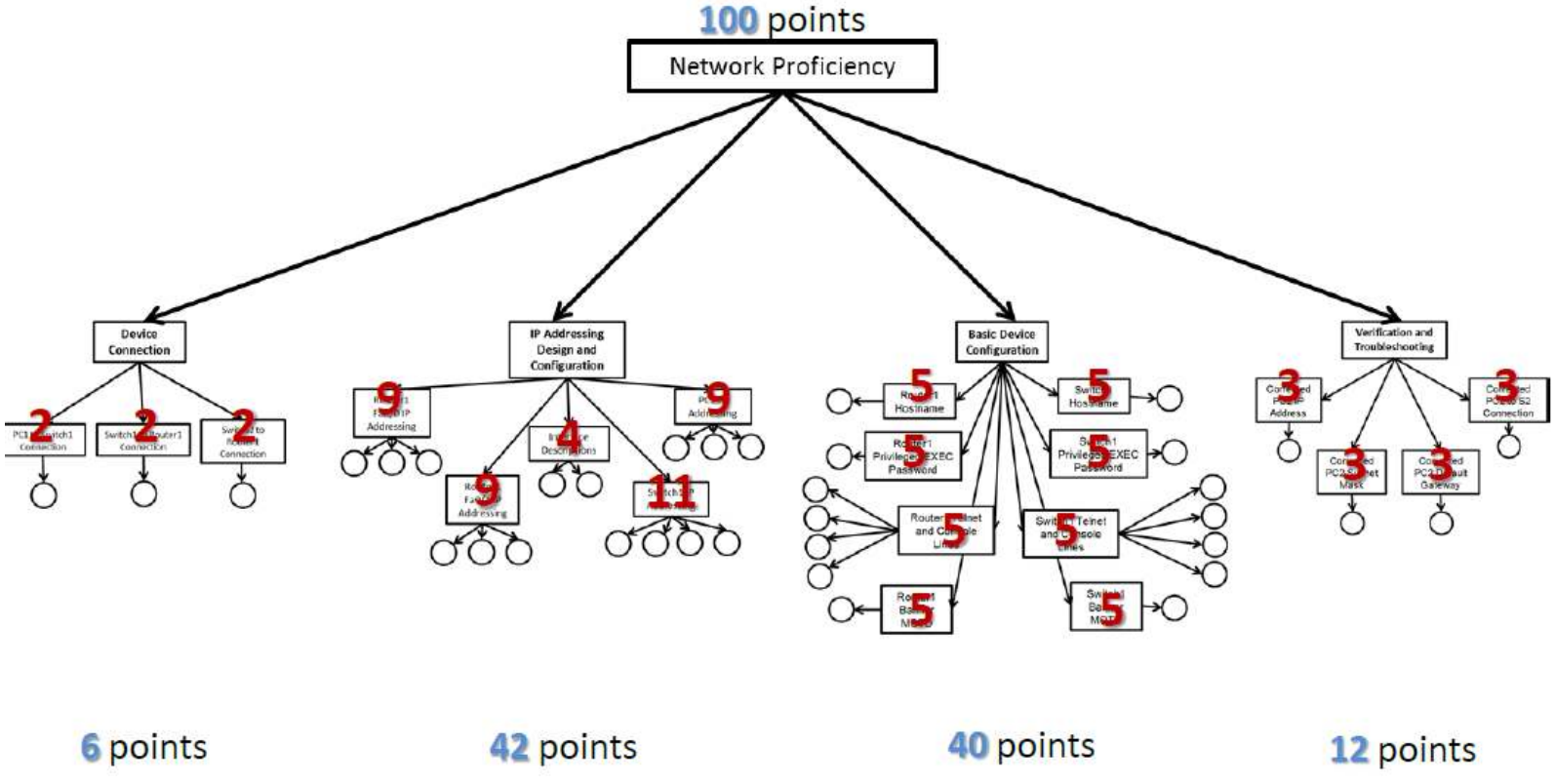

Fig. A1a. Graphical representation of the component-based scoring structure for the E1 PTSBA.

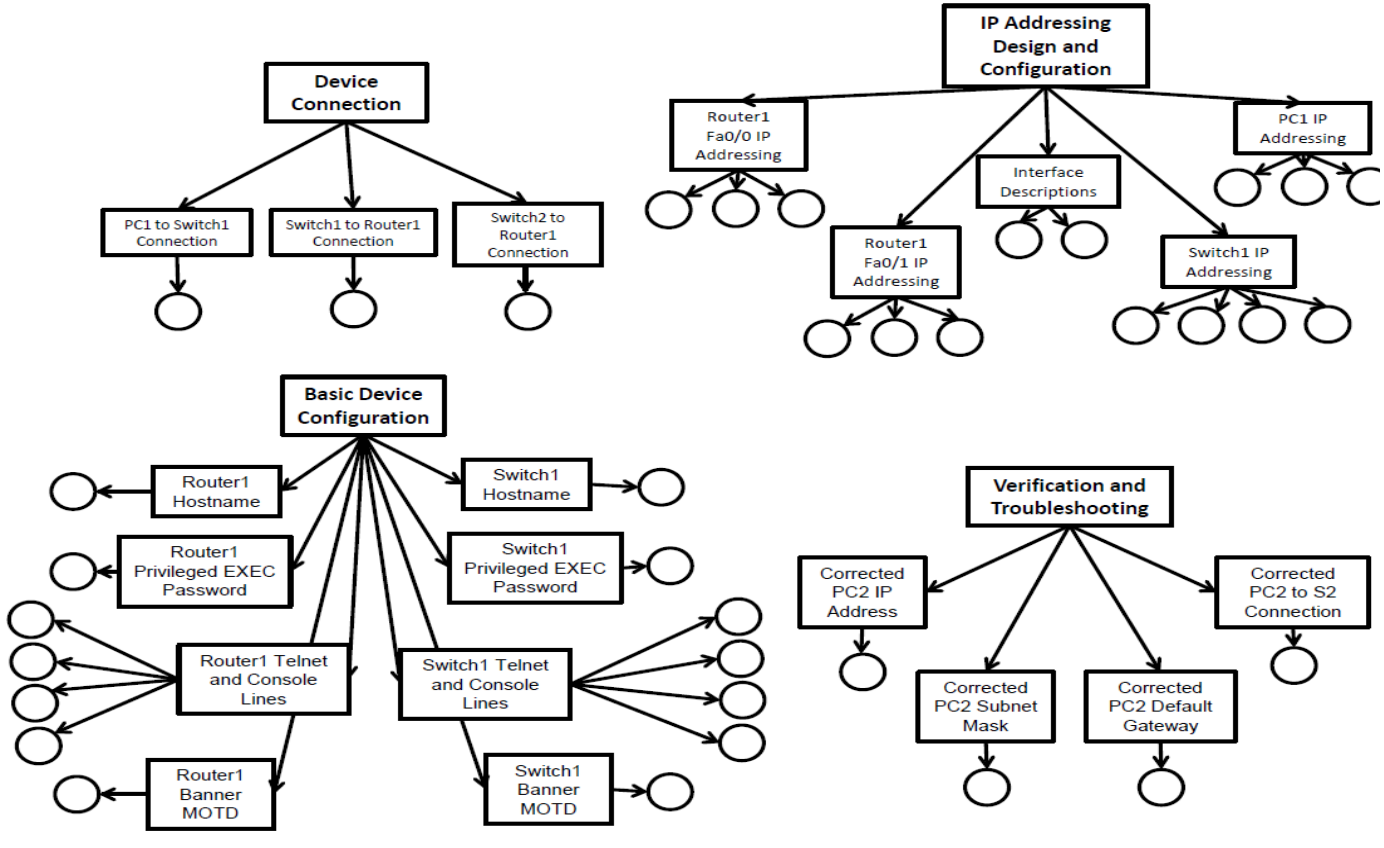

Fig. A1b. Graphical representation of the component-based scoring structure for the E1 PTSBA.

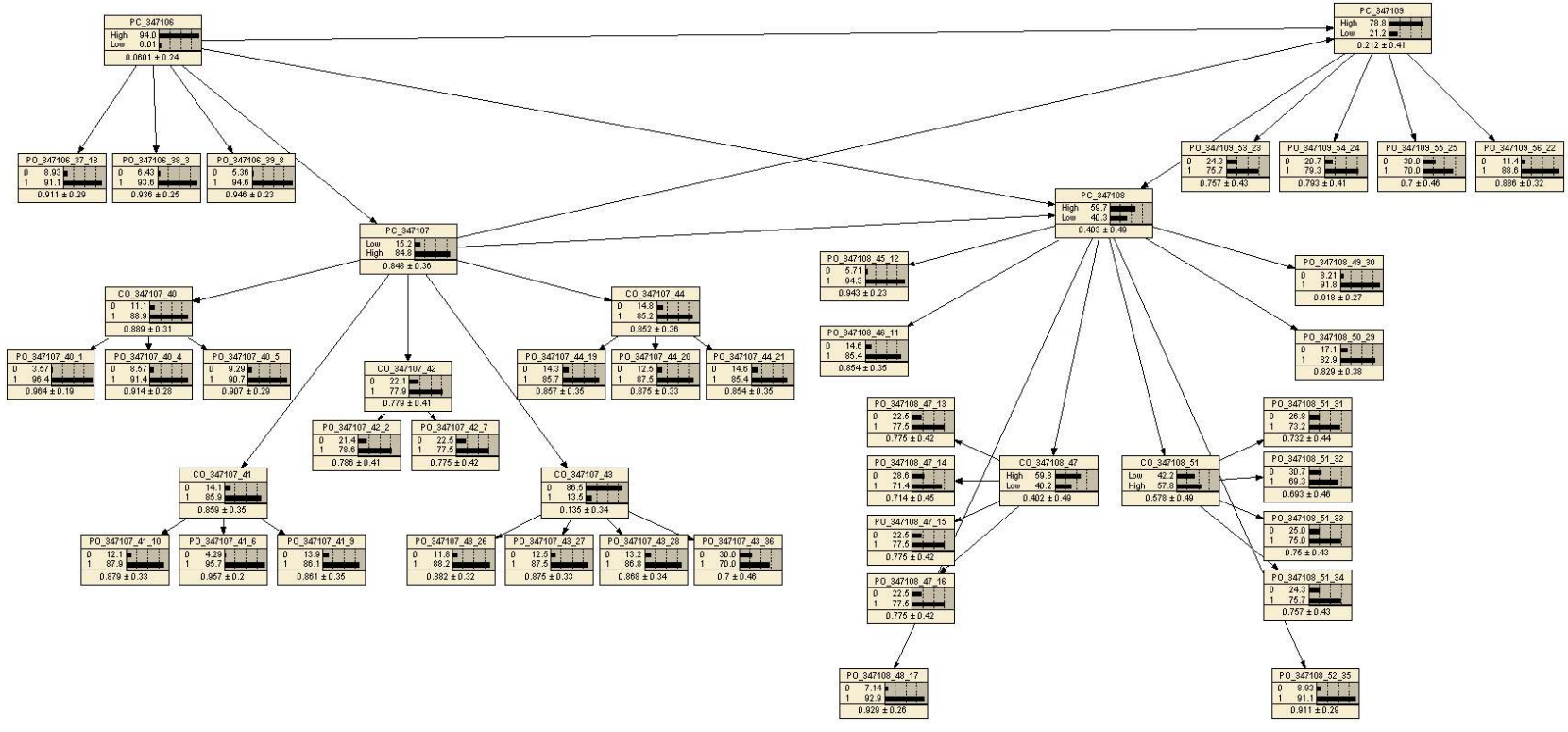

Fig. A2. Three-level Bayes net for the components-based scoring structure.

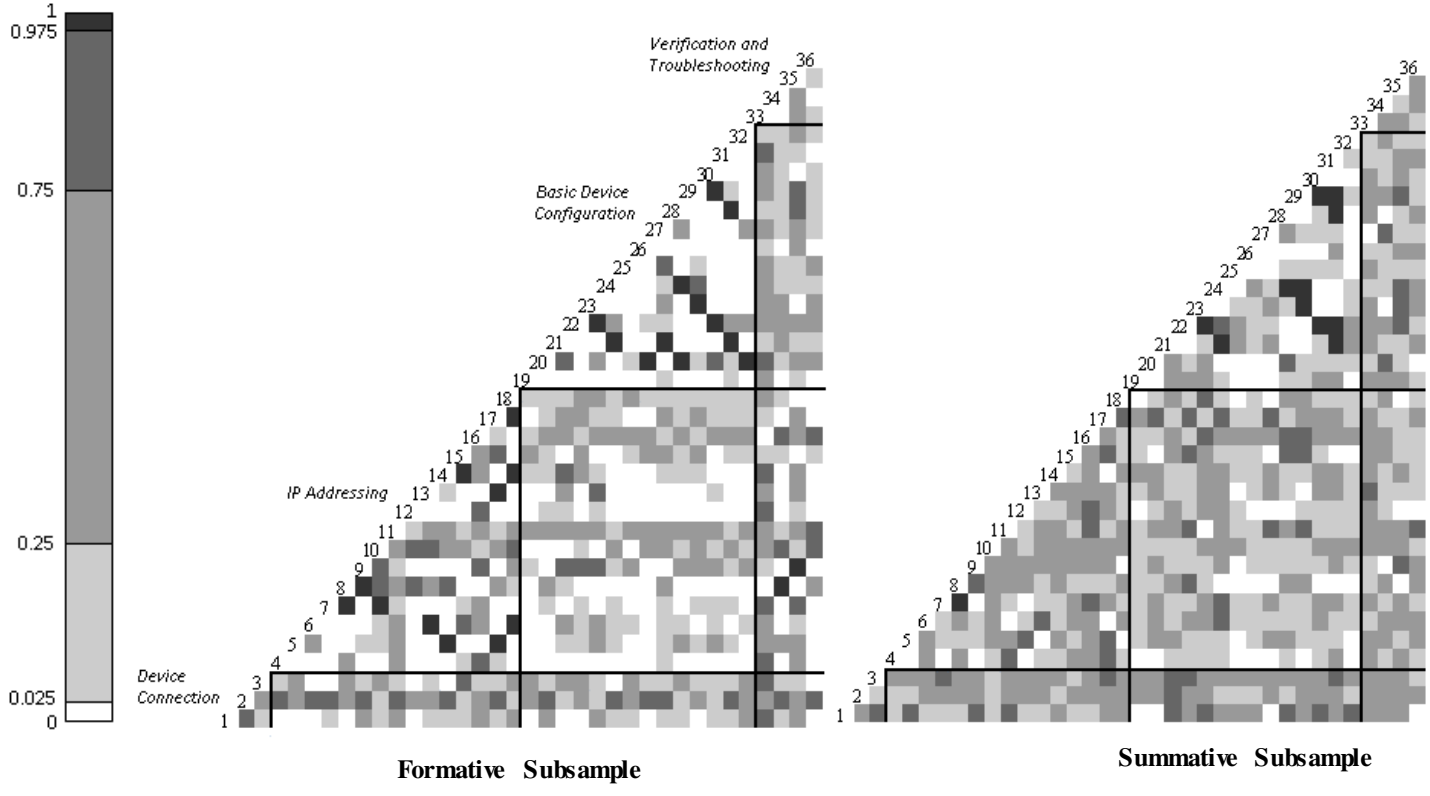

Fig. A3. Heat map for the Q3 statistic for pairs of score variables for both learner subsamples

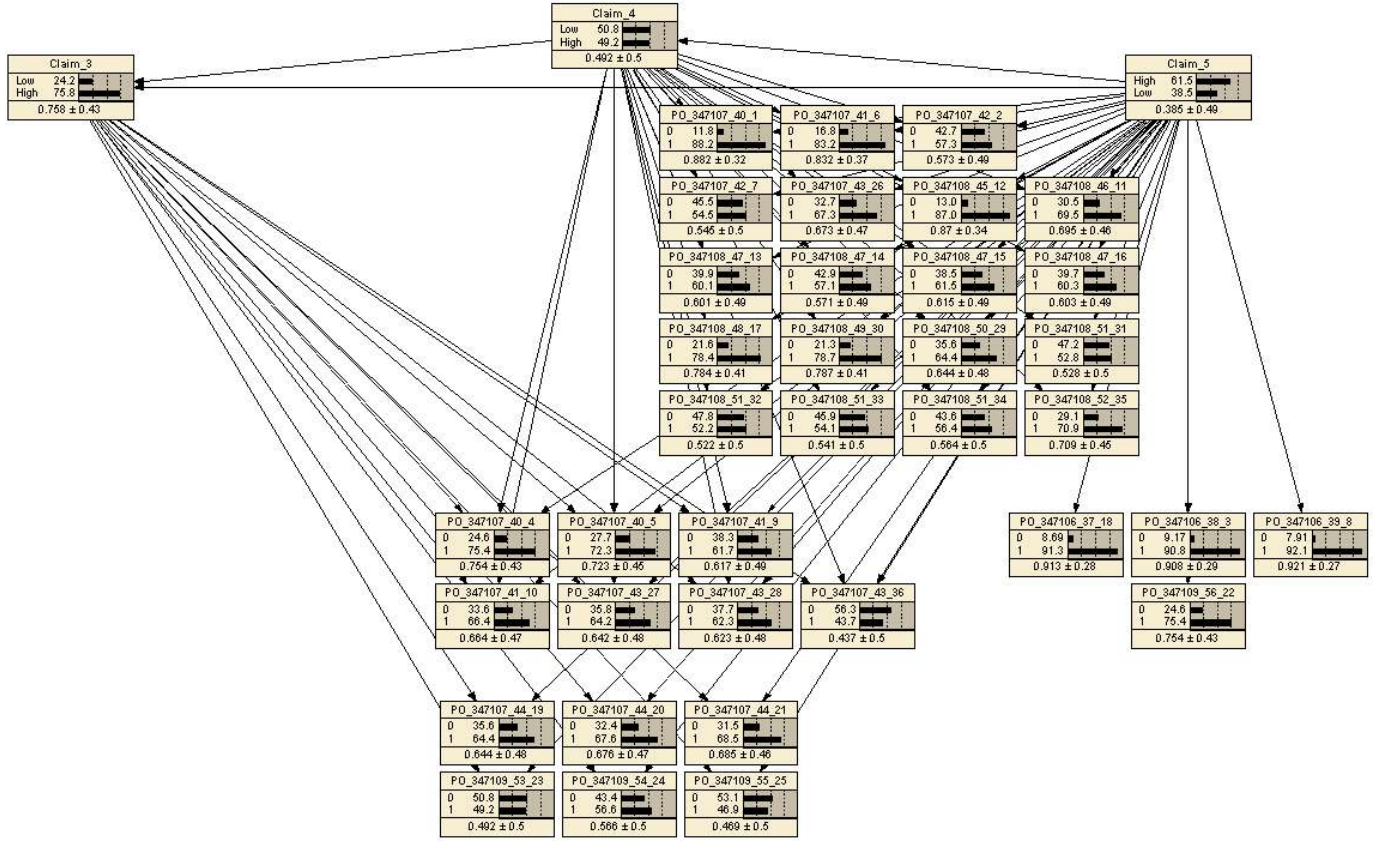

Fig. A4. Full Bayes net for the claims-based scoring structure.

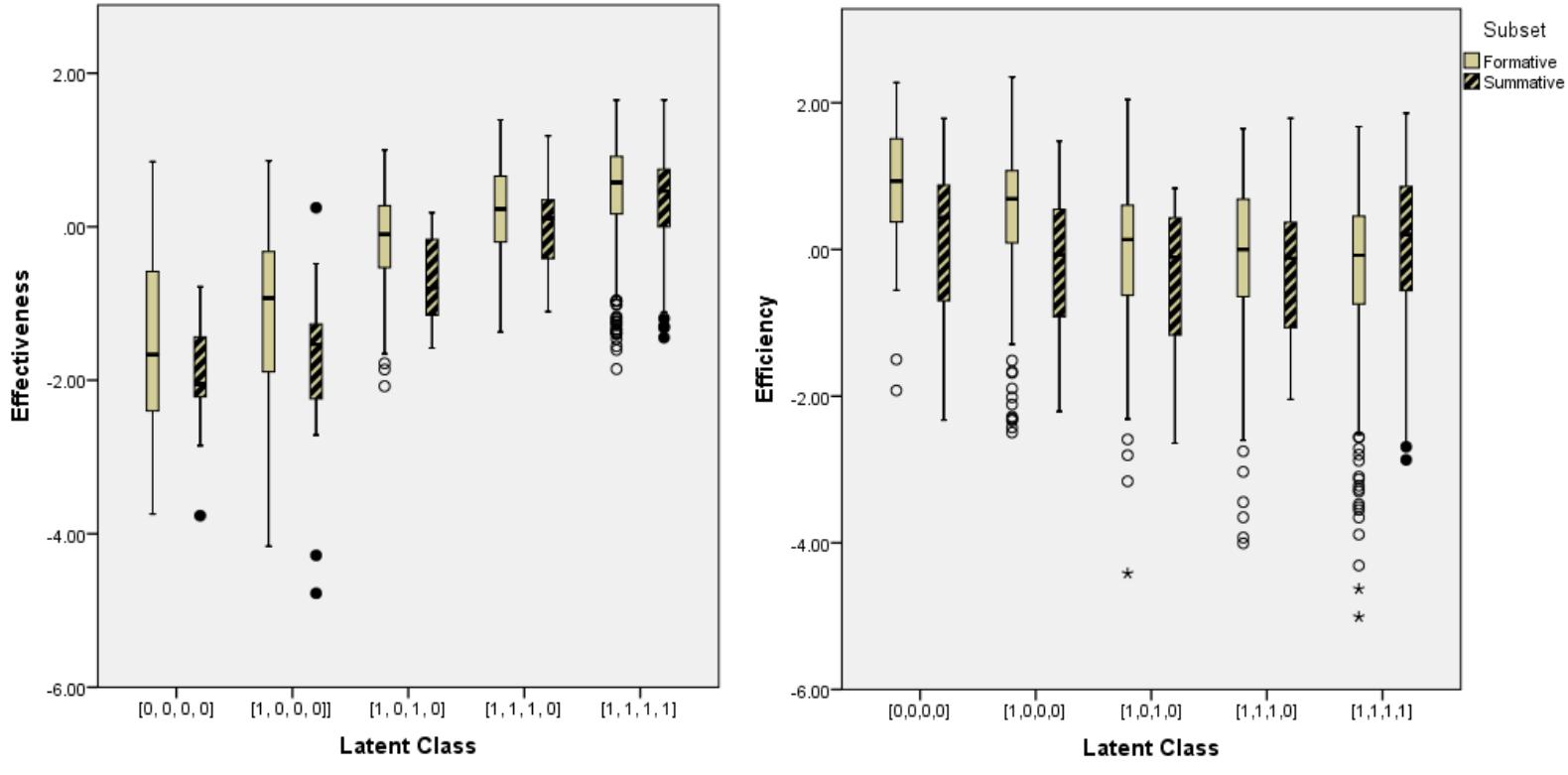

Fig. A5. Distributions of the effectiveness and efficiency scores across the five latent classes from the DCM with linear structural hierarchy for the component-based scoring structure.

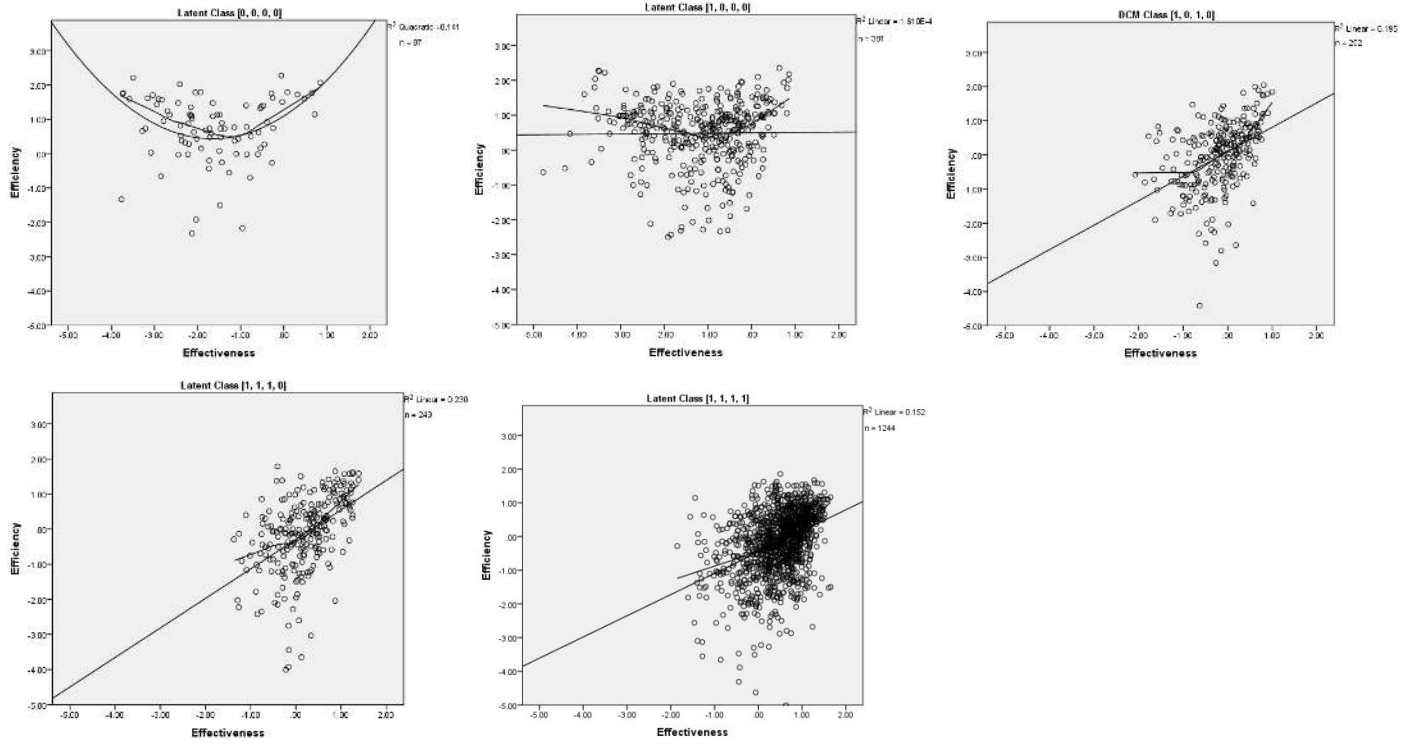

Fig. A6. Bivariate relations between the effectiveness and efficiency scores across the five latent classes in the DCM with linear structural hierarchy for the component-based scoring structure (parametric and nonparametric regression lines with R^2 and sample size values shown).