Evidence-centered Design for Diagnostic Assessment within Digital Learning Environments: Integrating Modern Psychometrics and Educational Data Mining

ANDRÉ A. RUPP
University of Maryland

REBECCA NUGENT
Carnegie Mellon University

and

BRIAN NELSON
Arizona State University

In recent years the educational community has increasingly embraced digital technologies for the purposes of developing alternative learning environments, providing diagnostic feedback, and fostering the development of so-called 21st-century skills. This special issue is dedicated to bridging recent work from the disciplines of educational and psychological assessment and educational data mining (EDM) via the assessment design and implementation framework of *evidence-centered design* (ECD). It consists of a series of five papers: one conceptual paper on ECD, three applied case studies that use ECD and EDM tools, and one simulation study that relies on ECD for its design and EDM for its implementation. In this introduction, we discuss the underlying rationales for the special issue in more detail, provide a short introduction to ECD, and describe the focus of the five selected papers.

Keywords: Educational data mining, psychometrics, digital learning environments, special issue.

Authors' addresses: André A. Rupp, Department of Human Development and Quantitative Methodology, University of Maryland, 1230-A Benjamin Building, College Park, 20742, ruppandr@umd.edu; Rebecca Nugent, Department of Statistics, Baker Hall 132, Carnegie Mellon University, Pittsburgh, PA, 15213, rnugent@stat.cmu.edu; Brian Nelson, School for Computing, Informatics, and Decision Systems Engineering, Arizona State University, Brickyard Engineering M1-04, Tempe, AZ, 85287, brian.nelson@asu.edu.

> [We propose] a vision of a world in which the natural instrumentation of a digital ocean blur[s] the distinctions between formative and summative assessment, curriculum and assessment, and formal and informal aspects of instruction. It is a world in which data are a side effect, not the primary goal of interesting and motivating activity, and perhaps a world where "testing" is a rare event, but assessment is "in the water". [DiCerbo and Behrens in press p. 302]

1. FROM A DIGITAL DESERT TO A DIGITAL OCEAN

As articulated eloquently in this quote and the article from which it was taken, the current ecosystem of educational assessment is increasingly moving from a formerly arid "digital desert" to a continually rising "digital ocean". Assessments within the digital desert are characterized by fixed-form stand-alone tests with selected or short response formats administered in paper-and-pencil format whereas assessment in the digital ocean is characterized by adaptive assessment suites that include performance-based tasks administered individually or collaboratively in digital learning environments.

Arguably, a move toward an integration of digital learning environments into educational learning contexts merely reflects the natural progression of continual and incidental learning within an increasingly digitally literate, facile, and connected society. This is particularly important when thinking of learning activities as both occurring within the K-12 school context and in various real-life situations. In such a world, assessment is much less separated from learning activities and becomes one of the many aspects of an integrated learning experience. Such a m omentum is already reflected in larger programmatic efforts by companies to restructure learning and assessment experiences as illustrated, for example, by the cognitively-based assessment of, for, and as learning initiative [Bennet 2010].

2. BECOMING METHODOLOGICALLY MULTILINGUAL

Motivated largely by these broader societal and disciplinary movements, specialists in different data-analytic traditions have to think about novel ways of applying and, most importantly, integrating different tools. In order to make sense of the "waves" of data within a particular learning context, researchers have to adapt traditional tools from classical test theory and generalizability theory [Brennan 2001; Crocker and Algina 2006; Raykov and Marcoulides 2011], modern psychometrics [de Ayala 2009; deBoeck and Wilson 2004], diagnostic measurement [Leighton and Gierl 2007; Rupp et al. 2010], multivariate statistics [Lattin et al. 2002], business process mining [van der Aalst 2011], and, of course, *educational data mining* (EDM) [Romero et al. 2010].

The boundaries among these disciplinary traditions are certainly not as clear as these labels might suggest. For example, all statistical techniques used in this list are inherently multivariate in nature, serve to condense or project data to lower-dimensional representations for the purposes of profiling or sorting them, and require stability and representativeness arguments for the results to be of use for practical decision-making. Moreover, successful projects that are situated in this multi-method space require interdisciplinary teams of researchers whose members have to become increasingly methodologically multilingual.

EDM techniques certainly appear to be very promising for addressing the challenges of automation of scoring procedures that can support meaningful feedback for learners, instructors, and other stakeholders. For example, the recent 2011 special issue of the journal *User Modeling and User-adapted Interaction* highlighted six different case studies that used a variety of EDM techniques to solve challenges of fine-tuning learning experiences within digital learning environments. It would seem that a blend of the best practices in educational and psychological measurement developed over the course of the last century coupled with these new data-analytic approaches is a fruitful starting point for leveraging the assessment as, of, and for learning opportunities that well-designed digital learning environments hold. In this special issue, we seek to illustrate how EDM techniques can be coupled successfully with principled assessment design to produce meaningful interpretations about learners.

3. THE EVIDENCE-CENTERED DESIGN FRAMEWORK

In order to facilitate the work in these interdisciplinary settings for educational assessment purposes, frameworks for the principled design and implementation of, as well as the data analyses and learner profile creation for, digital learning environments are needed. We chose the framework of *evidence-centered design* (ECD) for this special issue because of its increasing conceptual and practical traction in various assessment communities since the first foundational papers appeared about 10 years ago [e.g. Almond et al. 2001; Mislevy et al. 2003]. ECD has now been successfully applied to state large-scale science assessments [e.g. Zallas et al. 2010; see http://ecd.sri.com/], computer networking environments [e.g. Frezzo et al. 2009], and other assessment contexts. A special 2010 issue of the journal *Applied Measurement in Education* was dedicated to ECD and it is now increasingly taught in workshops at international conferences [e.g., Huff et al. 2012].

The specific organization of the components of the ECD framework into conceptual layers is described slightly differently by different authors. However, most refer to the five layers recently re-articulated by Mislevy [2011], which are *domain analysis*, *domain modeling*, the *conceptual assessment framework*, *assessment delivery*, and *assessment implementation*. From a practitioner's perspective, we find that these layers can be collected into two separate macro-layers: one layer of conceptual specifications and another layer of practical implementations of these specifications using digital technologies [see Mislevy et al. this issue].

3.1 The Conceptual Specification Macro-layer of ECD

Viewed from this distinction, the conceptual specification macro-layer of ECD consists of two key subcomponents. The first subcomponent represents the steps of *domain analysis* and *domain modeling* where researchers map out and formalize the key aspects of knowledge and performance that drive performance in a discipline or context of interest. The second subcomponent is driven by the *conceptual assessment framework* (CAF), which consists of three conceptual modules / models [see Mislevy et al., this issue, Figures 1 and 2].

In particular, *student models* are used to specify key elements of student proficiency about which one wants to infer, *task models* or *activity models* are used to specify the key features of tasks or activities that elicit performance driven by those proficiencies, and *evidence models* are used to specify the key statistical tools that are used to identify and accumulate evidence from observable performance into summary variables that relate to these proficiencies. Additionally, *assembly models* and *presentation models* can be used to specify the assembly and presentation of different activities to support the creation of a coherent evidentiary assessment argument.

3.2 The Practical Implementation Macro-layer of ECD

The practical implementation macro-layer of ECD is also known as the *four-process delivery architecture / four-process model* and consists of data management processes that map onto components in the conceptual specification layer [see Mislevy et al., this issue, Figure 3]. In this macro-layer, tasks / activities, scoring rules, and data management systems are being programmed, learners are interacting with the digital learning environments, and data are being stored and analyzed. In other words, the conceptual specifications of the first macro-layer are implemented in practice, which typically leads to an iterative cycle of conceptual and implementation refinements across both macro-layers.

The key processes in the practical implementation macro-layer are *activity selection*, *activity presentation*, *evidence identification*, and *evidence accumulation*. The first two relate to the *assembly model* and *presentation models* whereas the last two clearly relate to the *evidence models*, which conceptually connect the *student models* and *activity models*. In the evidence identification process, programmed scoring rules convert features of observable work products or log files of actions into *observable score variables*. In the evidence accumulation process, these score variables are then aggregated / synthesized (i.e. combined) using statistical *measurement models* from various disciplines to create higher-order summaries of learner performance.

4. OBJECTIVES OF THE SPECIAL ISSUE

To understand our objectives of this special issue, it is probably helpful to understand a few philosophical positions that we hold that are reflected in its composition. First and foremost, we do not believe that there is one single “true” – and often parametric – statistical model that is the “only” legitimate means of capturing key features of product and process data to characterize learners and activities. Rather, we believe that different processes of evidence identification and accumulation can provide competing, and jointly illuminating, perspectives on data from digital learning environments. As a result, we believe that utilizing a variety of statistical tools from the disciplinary traditions identified earlier, along with various strategies for model-data fit and inferential robustness, in the service of a coherent multi-lens understanding about data patterns is most promising.

Second, we firmly believe that such work benefits from the use of a principled assessment design framework. In this regard, we believe that ECD is the most promising, albeit not the only, candidate in this respect [see e.g. the assessment engineering principles reviewed by Zhou 2009]. We further believe that a coherent evidence-based understanding of student learning can only be possible when interdisciplinary teams of researchers continually work together within a design-based research approach using a framework such as ECD. Thus, we aim to promote open conversations about the often unspoken beliefs, assumptions, and values that researchers working to create coherent evidentiary assessment arguments within digital learning environments bring to the table. We believe that the ECD framework can be a powerful tool for structuring such discourse as we hope to promote with this special issue.

Consequently, we did not actively seek out projects that promote activities aimed at developing a fine-tuned machinery that can only support relatively simple summaries of student performance against already well-understood indicators of statistical quality (e.g., a computer-adaptive standardized test that yields a single global proficiency score). While such activities are clearly valuable, other publication venues are probably better suited to their dissemination. Rather, we sought to promote the principled exploration of methodological approaches from different disciplinary traditions that support more complex characterizations of student performance using appropriately adapted indicators of statistical quality and instructional utility. We believe that a continual cycle of a priori assessment design, in situ data exploration, and a posteriori theory refinement where data structures are messy, non-conventional ideas are embraced, and imperfect solutions are accepted, discussed, and used to inform future practice.

Rather than aiming for a comprehensive overview of the current status of the field, impossible within the confines of a single special issue and the staggered development of work in real life projects, we chose to highlight a few key projects that serve as illustrative case studies. Perhaps not surprisingly, the case studies in this issue do not necessarily present linearly organized stories about the added value of ECD for aligning evidence about learner development gathered through multiple methods. Rather, many authors describe both challenges and lessons learned from their work and acknowledge imperfections, road blocks, and open questions. We believe that this presentation of results is a much more honest reflection of the complexity of these endeavors. The power of ECD lies in unearthing the deep-structure communality of the challenges and associated solutions across projects despite the seemingly surface-structure idiosyncratic facets of each.

To be included as an illustrative case study in this special issue, an article needed to demonstrate an arguably added value of the use of the ECD framework and integrated data-analytic approaches for the creation of evidentiary assessment arguments. We also required that data analyses contained as much detail as possible about model-data fit assessment and other forms of robustness analyses. Finally, we asked authors to discuss the connections between their design choices and their steps and interpretations from data analysis from an ECD perspective. Each of the four selected contributions addresses at least several of these requirements in ways that we hope are insightful to the readers.

5. STRUCTURE OF THE SPECIAL ISSUE

The first article by Robert Mislevy, John Behrens, Kristen DiCerbo, and Roy Levy represents a thoughtful reflection upon the different disciplinary traditions that guide more “traditional” large-scale assessment work in the digital desert and more “innovative” diagnostic assessment within digital learning environments in the digital ocean. Both of these assessment traditions are prototypically associated with different statistical approaches, in particular modern latent-variable models for large-scale assessments and EDM tools for digital learning environments. The authors discuss how the use of the ECD framework, which provides a common language, rhetorical skeleton, and practical toolbox for evidentiary assessment arguments across disciplinary traditions can help colleagues trained under either tradition understand how they can learn from one another. Put in a nutshell, the authors use their practical experience with different types of assessment projects to argue that thoughtful a priori design plans are necessary to allow for rigorous a posteriori data analyses within an iterative cycle of hypothesis generation and refinement about how students learn in digital learning environments. An integrative set of statistical tools from modern psychometrics and EDM then serve as the rhetorical grammar for constructing coherent evidentiary narratives about cross-sectional learning snapshots and longitudinal learning traces.

The second article by André A. Rupp, Roy Levy, Kristen DiCerbo, Shauna Sweet, Aaron Crawford, Tiago Caliço, Martin Benson, Derek Fay, Katie Kunze, Robert Mislevy, and John Behrens is the first in a series of three applied case studies. In it, the authors describe how ECD was used to design a digital learning environment for network engineering skills, *Packet Tracer*, which is used worldwide by learners across a wide range of ages. They then present results from statistical analyses focused on the “product data” from this environment (i.e., scored configurations of networks), as well as the “process data” (i.e., log file entries). Importantly, their analyses span across methodological disciplines and include tools from classical test theory, traditional psychometric models, modern psychometric models, and multivariate data analysis / EDM methods. Through a series of model-fitting strategies across frameworks they show how the use of diverse tools provides evidentiary convergence. Moreover, they illuminate how it is possible to develop a more coherent understanding about student performance on the basis of product and process data. Throughout this work, they show how a constant reflection on the specifications of targeted assessment narratives via the ECD framework, coupled with constant interdisciplinary discussions within a diverse expert team, is necessary to yield evidentiary coherence. It is this evidentiary coherence that can lead to practically meaningful interpretations from the perspective of the developers of the learning environment.

The third article by Janice Gobert, Michael Sao Pedro, Ryan Baker, Ermal Toto, and Orlando Montalvo describes how tools such as text replay tagging and mining of the resulting tagged log files within a digital learning environment for science inquiry, *Science Assistments*, can help to assess learners’ current knowledge states as well as to provide targeted scaffolding. The authors discuss the two data-analytic stages in this work, namely (1) the development of detectors for the automated coding of log files and (2) the use of these tagged log files, along with other performance information, for deriving proficiency estimates. The article underscores nicely how activity constraints built into the system help make sequences of log files more interpretable than in relatively unconstrained systems. It also highlights that an effective application of EDM techniques consists of an adaptation of routines for automated data processing along with the specification of theoretical models of behavior and how these are manifested in data streams.

The fourth article by Deirdre Kerr and Gregory Chung illustrates how ECD was used to design a digital learning environment for basic mathematical skills in K-8 populations, *Save Patch*, and how the design choices informed cluster analyses of log files to understand strategy and performance differences of learners. The authors illustrate how the design of the learning environment, which here is segmented into distinct levels with relatively bounded tasks, reduces the complexity of the data-analytic endeavors to some degree. They compare the performance of different clustering algorithms for detecting learners who seem to be using different strategies characterized by different error profiles. The authors describe how iterative design, implementation, and data-analysis cycles were necessary to fine-tune the performance of the clustering routines and the associated interpretations.

The fifth article by Shauna Sweet and André A. Rupp represents a unique perspective on how ECD can be used for the design and analysis of simulation studies for understanding learner behavior within digital learning environments. Their simulation study was motivated by a digital learning environment, the epistemic game *Land Science*, which helps learners how to think, act, and communicate like professionals in professional disciplines based on mentorship experiences. Their case study helps illuminate the manifold challenges that researchers face when they want to learn about the robustness of novel non-parametric statistical methods. In particular, in the absence of well-specified parametric statistical models that could be used for data generation, researchers have to choose mechanisms for data-generation that mimic the key behaviors of learners while serving as well-targeted statistically designed experiments. The authors show how the use of the ECD framework can unify the design, implementation, and reporting of such simulation studies as the glue that binds the methodological inquiries and the real-life design activities for the digital learning environment together.

In a final synthesis piece, we reflect upon some of the key lessons that we have learned from reading and editing these contributions. We specifically reflect on the key issues and themes that we noted throughout the articles in terms of how the ECD framework was used, how different statistical methods were used to support evidentiary narratives with respect to the framework, and how the use of ECD and modern data-analytic tools served instructional decision-making purposes.

6. A LOOK AHEAD

We certainly have enjoyed engaging in the process of creating this special issue and hope that you will similarly enjoy reading the contributions. As is typically the case, many additional projects are currently being developed whose dissemination would similarly enrich our understanding of the key ideas brought to the forefront. We thus see this special issue as a contribution to the ongoing discussion and exploration of lessons learned in the complex endeavor of creating evidence-based diagnostic assessment narratives within digital learning environments, rather than as a definitive collection of some sort.

To continue the conversation, we are in the process of creating a website entitled "Digital ECD" (www.digitalecd.org) – a non-public prototype already exists – where we invite you to submit a snapshot of the evidentiary reasoning process for your project. If you are interested in doing so, please contact André A. Rupp (ruppandr@umd.edu) – we look forward to hearing from you and hope that you enjoy reading this special issue!

Sincerely,

André A. Rupp, Rebecca Nugent, and Brian Nelson
(Co-editors, JEDM Special Issue)

REFERENCES

APPLIED MEASUREMENT IN EDUCATION. 2010. *Evidence-centered assessment design in practice* (special issue), 23(4).

ALMOND, R. G., STEINBERG, L. S., & MISLEVY, R. J. 2001. *A Sample Assessment Using the Four Process Framework.* (CSE Tech. Rep. 543). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

BENNET, R. 2010. Cognitively based assessment as, of, and for learning (C-BAL): A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8*(2-3), 70-91.

BRENNAN, R. J. 2001. *Generalizability theory*. New York, NY: Springer.

CROCKER, L., & ALGINA, J. 2006, 2nd ed.. *Introduction to Classical and Modern Test Theory*. Wadsworth, Belmont, CA.

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

DEBOECK, P., & WILSON, M. 2004. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer, New York, NY.

DICERBO, K. E. & BEHRENS, J. T. 2012. Implications of the digital ocean on current and future assessment. In R. Lissitz, Ed., *Computers and their Impact on State Assessment: Recent History and Predictions for the Future* (pp. 273-306). Information Age Publishing, Charlotte, NC.

FREZZO, D.C., BEHRENS, J.T., & MISLEVY, R.J. 2009. Design patterns for learning and assessment: facilitating the introduction of a complex simulation-based learning environment into a community of instructors. *The Journal of Science Education and Technology.* Springer Open Access http://www.springerlink.com/content/566p6g4307405346/

HUFF, K., EWING, M., HENDRICKSON, A., KALISKI, P., & PACKMAN, S. 2012, April. *Applications of Evidence-centered Design (ECD) in Large-scale Assessment.* Workshop given at the annual meeting of the National Council on Measurement in Education (NCME), Vancouver, BC, Canada.

LATTIN, J., CARROLL, D., & GREEN, P. 2002. *Analyzing Multivariate Data*. New York, NY: Duxbury Press.

LEIGHTON, J., & GIERL, M. 2007. Eds. *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge University Press, New York, NY.

MISLEVY, R. J. 2011. *Evidence-centered design for simulation based assessment* (CRESST Report 800). University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Los Angeles, CA.

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

RAYKOV, T., & MARCOULIDES, G. 2011. *Introduction to psychometric theory*. Routledge, New York, NY.

ROMERO, C., VENTURA, S., PECHENIZKIY, M., & BAKER, R. S. J. D. 2010. Eds. *Handbook of educational data mining*. Chapman & Hall / CRC, New York, NY.

RUPP, A. A., TEMPLIN, J., & HENSON, R. A. 2010. *Diagnostic measurement: Theory, methods, and applications*. Guilford Press, New York, NY.

USER MODELING AND USER-ADAPTED INTERACTION. 2011, special issue. *Data mining for personalised educational systems*, 21(1-2).

VAN DER AALST, W. M. P. 2011. *Process mining: Discovery, conformance and enhancement of business processes*. Springer, New York, NY.

ZALLES, D., HAERTEL, G., & MISLEVY, R. 2010. *Using Evidence-Centered Design to Support Assessment, Design and Validation of Learning Progressions (Large-Scale Assessment Technical Report 10).* SRI International, Menlo Park, CA.

ZHOU, J. 2009. *A review of assessment engineering principles with select applications to the Certified Public Accountant examination* (Technical Report W0903). American Institute of Certified Public Accountants, Inc. Available online via http://www.aicpa.org/Pages/Default.aspx