JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

Journal of Educational Measurement
Spring 2017, Vol. 54, No. 1, pp. 3–11

Computational Psychometrics in Support of Collaborative Educational Assessments

Alina A. von Davier
ACTNext, by ACT, Inc.

The rapidly growing literature on teamwork and collaborative problem solving suggests that these skills are becoming increasingly important in today's education and workforce. This special issue of *Journal of Educational Measurement* seeks to channel the contributors' expertise toward advances in the measurement and assessment of cognitive and noncognitive skills of individuals and teams in education. The articles in this special issue discuss the challenges and opportunities for developing collaborative assessments that cover the assessment of collaborative skills, of collaborative problem solving skills, and/or the assessment of cognitive skills through collaborative problem solving tasks. These articles combine collaborative assessment methods with advanced psychometrics techniques, such as computational psychometrics, for analyzing collaborative behavior in educational assessments; they introduce the recent progress on innovative ways of studying collaboration in education.

Recent developments indicate that society is interested in redesigning educational assessments and not merely improving the assessments we have. Educators request assessments that reflect the way people actually teach, learn, and work. There is a renewed interest in performance assessments and efforts are being made to develop these complex assessments in virtual settings. One type of performance task is the collaborative task, where an individual works together with one or more agents to solve a problem. This ubiquitous interactive setting of our everyday life poses challenges for assessment. Questions about which skills are needed in such a context, how to disentangle the individual contributions from the team contribution, and what types of models could help us predict a successful performance have tempered the enthusiasm around pursuing the development of educational assessments with collaborative components. Nevertheless, other recent events indicate that progress is being made. For example, the Programme for International Student Assessment (PISA) 2015 administered a test of collaborative skills (OECD, 2013); the National Assessment of Educational Progress (NAEP) hosted a symposium on collaborative problem solving (CPS) in September, 2014, and as a follow-up, the National Council on Measurement in Education (NCES) commissioned a white paper on the considerations for introduction of the CPS in NAEP; the College Board's Advanced Placement Computer Science Assessment is being redesigned and one of the new features is introducing collaborative tasks; Educational Testing Service (ETS) and the Army Research Institute co-hosted a working meeting, Innovative Assessment of Collaboration, November 3–4, 2014, and an edited interdisciplinary volume based on that meeting is being published with Springer Verlag (von Davier, Zhu, & Kyllonen, 2017); the Smarter Balance Consortium developed an assessment system where performance tasks, including collaborative tasks, are considered for being administered to students as a preparatory experience and then are followed with individual assessment (see Davey et al., 2015, Chapter 4).

In this introductory article I focus on three measurement-related aspects of the process of building collaborative assessments. First, I briefly discuss the data types collected from the collaborative assessments and the dependencies in the data that force psychometricians to rethink the approach to measurement of these complex constructs. This will lead directly to my second point in this article: the introduction of a recent discipline, *computational psychometrics*, as a blend of stochastic processes theory, computer science–based methods, and theory-based psychometric approaches that may aid the analyses of complex data from performance assessments, including collaborative assessments. This discipline organically developed around the complex next-generation learning and assessment systems that include performance tasks, such as collaboration, games, and simulations. Third, I briefly describe a statistical model for collaborative activities. In this introduction I will not discuss the constructs of collaboration or collaborative problem solving *per se*, nor the infrastructure challenges for administering collaborative tasks and for the collection of rich data, nor the validity of these assessments. Each of the subsequent articles will address these issues in the context of their applications. This introductory article will conclude with an overview of the rest of the articles in this special issue.

Clearly, developing collaborative assessments is challenging and requires an interdisciplinary approach. This collection of articles illustrates these challenges in various applications, including the study of measurement invariance in collaborative tasks.

Data, Log Files, and Data Dependencies

Collaborative tasks are interactive. This means that test takers talk, negotiate, hypothesize, revise, and respond, orally, with gestures, and online with chats and emoticons, acronyms, and so on. All of these data are *process* data that offer an insight into the interactional dynamics of the team members; they are relevant for defining collaborative tasks and for evaluating the results of the collaboration (see also Agard & von Davier, in press). Traditionally, these data were not available to researchers at a scale that would allow for meaningful inferences. With the advances in technology, these complex data can be captured in computerized or in virtual settings.

The data from collaborative tasks consist of time-stamped sequences of events registered in a log file. From a statistical perspective, these activity logs or *log files* are detailed time series describing the actions and interactions of the users. (See also Hao, Smith, Mislevy, von Davier, and Bauer [2016] for a discussion and description of the log files for the collaborative assessments). In addition to the process data, if the collaboration is set up in a cognitive (say, science) task it will also result in *outcome* data. These types of data are more similar to the outcome data from the traditional tests and indicate if a particular question was answered correctly and whether the problem was solved (or to what degree it was solved).

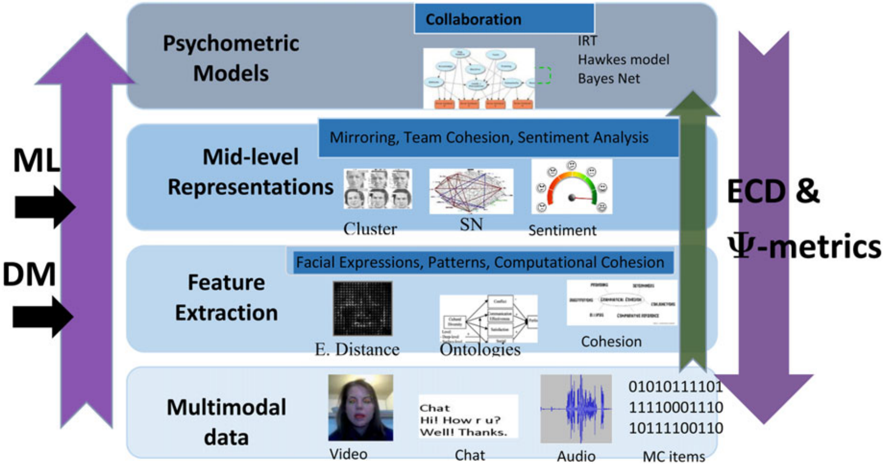

Figure 1. Computational psychometrics. A model for performance and collaborative assessments.

Another feature of the data from collaborative tasks is that they can be at different units of measurement: they can be characterized as individual and team data both as outcome data and process data. In order to save space, these types of data and their role in assessment are discussed further in the online Supporting Information.

Computational Psychometrics

As mentioned above, computational psychometrics is defined as a blend of data-driven computer science methods (machine learning and data mining, in particular), stochastic theory, and theory-driven psychometrics in order to measure latent abilities in real time (von Davier, 2015).

This blend can be instantiated as iterative and adaptive hierarchical inference data algorithms embedded in a theoretical psychometric framework as shown in Figure 1. This visual representation of the model is adapted from a figure from Khan (2015) on the multimodal hierarchical approach. This hierarchical approach to multimodal data has been discussed in Khan, Cheng, and Kumar (2013).

The idea illustrated in Figure 1 is that the approach to test development and data analysis is rooted in theory and starts with the implementation of evidence centered design (ECD; Mislevy, Steinberg, Almond, & Lukas, 2006) principles; then the test is piloted and the multimodal metadata (fine grain data points) are collected along with the data from traditional items. This approach is also known as a top-down approach because it relies on input and theories devised by human experts. The next step involves a bottom-up approach, in which the data are analyzed with unsupervised and supervised algorithms from data mining and machine learning. If new relevant patterns are identified (for example, performance patterns, behavior patterns), these will be considered for incorporation in the revised psychometric models. The data mining (DM) and machine learning (ML) approaches applied to metadata will result in midlevel representations of the constructs that can be further incorporated in the original psychometric models along with the data from the traditional items.

Next, the psychometric models are updated and the process is repeated with a second large-sample pilot data. At this stage, one may consider models for stochastic processes if the data allow. Once the psychometric model is stable, only then is the assessment administered to the population of interest. On the operational data, only supervised machine learning algorithms and already defined and validated psychometric models are further used in order to achieve a stable measurement and classification rules. The steps and cycles depicted in Figure 1 are discussed next in more detail.

How to Instantiate the Computational Psychometrics (CP) Model From Figure 1?

The use of ECD in the CP context. This part is the easiest to understand for psychometricians. This framework involves designing the system (learning and/or assessment) based on theory, identifying constructs associated with competency of interest (*evidence*), and finding evidence for these constructs from low-level multiple sensory data (*evidence about evidence*; Khan, 2015). At this stage, one defines the constructs to be measured and *develops the tasks* that are collaborative so that they provide the "right" data to support the intended claims and choice of psychometric models. We may also consider embedding the CPS tasks into traditional assessments to increase the measurement accuracy for the cognitive construct and to enhance the data around the CPS task (Liu, Hao, von Davier, Kyllonen, & Zapata-Rivera, 2015).

The psychometric models for the CP context. The types of models associated with complex data with dependencies have been item response theory (IRT)-like models and Bayesian belief networks (BBNs; Levy, 2014: Mislevy et al., 2014). BBNs model the probability that a student has mastered a specific knowledge component conditional on the sequence of responses given to previous elements of a task, and eventually to other tasks, whether they are associated with that knowledge component or not (as long as they are part of the network and share at least an indirect connection). BBNs have long been applied in simulations and games to represent student knowledge and thereby guide the activities of the tutoring system (Corbett & Anderson, 1995; Desmarais & Baker, 2012; VanLehn, 2008). BBNs have also played a central role in the design of complex assessments (Shute, Hansen, & Almond, 2008); therefore, BBNs are an obvious methodological bridge between measuring CPS skills and traditional psychometric theory. However, the practical implementation of BBNs often requires highly simplifying assumptions, and, as with traditional models, they have not been adapted to represent the knowledge of multiple individuals simultaneously.

Fortunately, there are other models such as stochastic point processes that have been used extensively in economics that can aid the modeling of interdependencies based on the temporal structure of the collaborative interactions (von Davier & Halpin, 2013), hidden Markov models (see Soller & Stevens, 2008), and models rooted in the cognitive or social theories such as agent-based modeling, ACT-R (Bergner, Andrews, Zhu, & Kitchen, 2015) and Markov decision processes, which is a cognitive model with separable components (goals/motivation, beliefs about the world, ability to optimize behavior) and which defines behavior as an optimization of expected rewards based on current beliefs about the world (LaMar, 2014).

The application of DM and ML tools in the CP context. DM does not have a long history in education or psychology because, until recently, educational and psychological data were not often of high enough dimensionality to require such techniques. The purpose of data mining techniques is to reduce the dimensionality of the data set by extracting interpretable patterns to allow research questions to be addressed that would not otherwise be feasible (Romero, Gonzalez, Ventura, del Jesus, & Herrera, 2009). Different clusters (patterns of responses) may be assigned different scores. The tools known as visualization, clustering, classification, feature extraction, sequence clustering, and editing distance methods for scoring are examples of DM tools that may be applied. There are already promising results regarding the identification of new evidence to feedback into ECD/test development cycle and generation of testable hypotheses (see Kerr, 2015; Kerr & Chung, 2012; Zhang, Hao, Li, & Deane, 2015).

Machine learning algorithms may be used for *in vivo* adaptive "learning" and assessment, by using the results/features from a DM classifier (often in conjunction with a psychometric model in a Bayesian framework) that "learns" from the data to predict the success on a task. ML algorithms rely on the availability of large and representative training data sets. These algorithms have been used in education for the development of automated scoring of essays; now we are using similar algorithms for the automatic scoring of speech and chat in collaborative interactions and for the automated detection of affective states during the collaboration (see Khan, 2015; von Davier, van der Schaar, & Baraniuk, 2016; Wang, Hao, Liu, Chen & von Davier, 2015).

In specific practical applications, this hierarchical iterative framework may be implemented in simplified or less explicit forms; for example, some of the steps may be there but may not be explicitly described and some of the steps may not be needed.

A Statistical Model for Collaboration

One of the innovations in the measurement of collaborative problem skills was presented in von Davier and Halpin (2013). They modeled collaboration as statistical dependence among the activities of two or more individuals. They proposed to measure the degree of interdependence demonstrated by the activities of the individuals in a team by using the Kullback-Leibler divergence (KL) of the marginal distributions (of the event sequences of the actions of each individual) from the joint distribution (of the team). The interpretation of KL in the context of collaboration is intuitive. If KL $= 0$ then this is an independent team, in which the members are working independently, like a gymnastics team, for example (see von Davier & Halpin, 2013, for a discussion of different types of teams based on the degree of dependence). When KL is positive some interdependence is exhibited among the activities of the individuals, with larger values indicating more interdependence (i.e., a greater divergence from the model of independence).

Von Davier and Halpin (2013) defined an outcome as a function of the complete time series, or sequences of actions (of the process data from each team member). If the activities recorded are correct responses to the components of a CPS task, we could define the group's total score on the task as the sum of the outcomes from all team members to all task components and compute its expected value accounting for the degree of interdependence. For example, for an independent team, the expected outcome is simply the sum of its parts. A productive collaboration can then be defined as one for which the expected outcome is greater than the sum of its parts, and an unproductive collaboration would have an expected performance worse than that of the independence model. A similar approach can be applied to other collaborative outcomes. Instead of sum scores, it will be generally advantageous to have a psychometric model for the entire response pattern, for instance, a (modified) IRT model or a Bayesian network.

Hence, in modeling the *processes* of collaboration, we are concerned about describing the statistical dependence exhibited by the activities of groups of individuals. In modeling *the outcomes* of collaboration, we are concerned with judging the performance of a group relative to what we would expect from the individual group members had they not collaborated. In CPS we are concerned with how the probability of an individual's activities changes over continuous time as a function of previous activities. These aspects of CPS are discussed in this special issue.

The Structure of the Special Issue

The articles included in this issue focus on different parts of developing educational assessments of collaboration: they discuss the framework for the skills to be measured, the psychometric properties of the test, the data collection design, and the measurement models. The first two articles focus more on the task and test design, while the rest of the articles focus on methodological aspects of the CPS assessments. In some of these articles the computational psychometric model is more obvious than in others.

Scoular, Care, and Hesse's article "Designs for Operationalizing Collaborative Problem Solving for Automated Assessment" outlines general design principles of collaborative problem solving curriculum-embedded assessments; it examines the degree to which assessments of collaborative problem solving skills have the capacity to discriminate between the contributing subskills. This has implications for current theoretical frameworks for collaborative problem and for the teaching of the skills.

Rosen's article "Assessing Students in Human-to-Agent Settings to Inform Collaborative Problem-Solving Learning" explores challenges in the development and validation of such assessments. The design principles and validation processes are applied to an empirical data set based on the *Animalia* collaborative science problem solving international project from Pearson.

Andrews, Kerr, Mislevy, von Davier, Hao, and Liu's article "Modeling Collaborative Interaction Patterns in a Simulation-Based Task" describes the use of process data and performance outcomes in the *tetralogue* (a simulation-based collaborative science task at ETS) to examine gender and cultural differences in collaboration.

Analyses using an Andersen/Rasch multivariate model explore the propensities of particular dyads to interact in accordance with certain patterns of interaction.

Halpin, von Davier, Hao, and Liu's article, "Measuring Student Engagement During Collaboration" elaborates the use of point processes and related methods for modeling interdependence in multivariate time series data. Point processes model the dependence in timing among collaborators' actions. This provides an intuitive measure of engagement among collaborators and is used in this article to develop "collaboration indices." The approach was illustrated with the same data from the *tetralogue* as in Andrews et al.

In Wilson, Gochyyev, and Scalise article, "Modeling Data From Collaborative Assessments: Learning in Digital Interactive Social Networks" the focus is on the assessment of cognitive skills through collaborative tasks, using initial field-test results from the Assessment and Teaching of 21st Century Skills (ATC21S) project. Specifically, the project's "ICT literacy—Learning in digital networks" is investigated here. The article includes a description of the development of the learning progression. Modeling of results employs unidimensional and multidimensional item response models, with and without random effects for groups.

The article of Herborn, Mustafić, and Greiff, "Mapping an Experiment-Based Assessment of Collaborative Behavior Onto Collaborative Problem Solving in PISA 2015: A Cluster Analysis Approach for Collaborator Profiles," conceptually embeds a computer-based human-agent collaborative behavior assessment (COLBAS) into the PISA 2015 CPS approach. In this article a model-based cluster analysis is presented; the model was employed to identify profiles of collaborators.

Olsen, Aleven, and Rummel's article, "Statistically Modeling Individual Students' Learning Over Successive Collaborative Practice Opportunities," presents an extension of the additive factors model (AFM) used in the educational data mining community; this is a standard logistic regression model for modeling individual learning, often used in conjunction with knowledge component models and tutor log data. The extended model predicts performance of students solving problems collaboratively with an intelligent tutoring system.

In conclusion, although there is agreement that collaboration is an important set of skills (Griffin & Care, 2015), there is less agreement on how to build an accurate assessment at scale to measure those skills. This special issue is a first attempt to describe several pioneering measurement approaches and applications in educational measurement.

Acknowledgments

This work was completed while the author was employed at Educational Testing Service. The author thanks Jim Carlson, Randy Bennett, Isaac Bejar, and Qiwei He for their suggestions on a previous version of this paper. The opinions expressed in this paper are those of the author and not necessarily of Educational Testing Service or ACT.

References

Agard, C., & von Davier, A. A. (in press). The virtual world and the reality of testing: Building virtual assessments. In H. Jiao & R. Lissitz (Eds.), *Technology-enhanced innovative* *assessment: Development, modeling, and scoring from an interdisciplinary perspective*. Charlotte, NC: Information Age.

Bergner, Y., Andrews, J. J., Zhu, M., & Kitchen, C. (2015, July). *Agent-based modeling of collaborative problem solving*. Paper presented at the 10th annual Interdisciplinary Network for Group Research (INGRoup) conference, Pittsburgh, PA.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*, 253–278.

Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N., & Wise, L. (2015). *Psychometric considerations for the next generation of performance assessment*. Princeton, NJ: Educational Testing Service.

Desmarais, M. C., & Baker, R. S. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User Adaptive Interaction*, *22*, 9–38.

Griffin, P., & Care, E. (2015). *Assessment and teaching of 21st century skills*. Berlin, Germany: Springer Verlag.

Hao, J., Smith L., Mislevy, R., von Davier, A., & Bauer, M. (2016). *Taming log files from game and simulation based assessment: Data model and data analysis tool* (Research Report 16–10). Princeton, NJ: Educational Testing Service.

Kerr, D. (2015). Using data mining results to improve educational video game design. *Journal of Educational Data Mining*, *7*(3), 1–17. Retrieved December 27, 2016, from http://www.educationaldatamining.org/JEDM/index.php/JEDM/article/view/JEDM048.

Kerr, D., & Chung, G. K. W. K. (2012). Identifying key features of student performance in educational video games and simulations through cluster analysis. *Journal of Educational Data Mining*, *4*, 144–182.

Khan, S. (2015, September). *Multimodal behavioral analytics for intelligent training and assessment systems*. Seminar presentation at the Rutgers University School of Engineering. Retrieved December 27, 2016, from http://www.ece.rutgers.edu/node/2111

Khan, S., Cheng, H., & Kumar, R. (2013). A hierarchical behavior analysis approach for automated trainee performance evaluation in training ranges. In D. D. Schmorrow & C. D. Fidopiasti (Eds.), *Foundations of augmented cognition*. Proceedings of HCI International 2013, July 2013, Las Vegas, NV. Heidelberg, Germany: Springer Verlag. Retrieved December 27, 2016, from https://books.google.com/books?id=jpa5BQAAQBAJ&pg=PA68&dq=khan,+cheng+and+kumar&hl=en&sa=X&ved=0ahUKEwito4jLr4rNAhWBWD4KHQ2KCNsQ6AEIHTAA#v=onepage&q=khan%2C%20cheng%20and%20kumar&f=false.

LaMar, M. M. (2014). *Models for understanding student thinking using data from complex computerized science tasks* (Doctoral dissertation). University of California, Berkeley. Retrieved December 27, 2016, from http://gradworks.umi.com/36/86/3686374.html

Levy, R. (2014). *Dynamic Bayesian network modeling of game-based diagnostic assessments (*CRESST Report 8037*)*. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Liu, L., Hao, J., von Davier, A. A., Kyllonen, P., & Zapata-Rivera, J.-D. (2015). A tough nut to crack: Measuring collaborative problem solving. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on computational tools for real-world skill development*. Hershey, PA: IGI-Global.

Mislevy, R. J., Corrigan, S., Oranje, A., Dicerbo, K., John, M., Bauer, M. I., Hoffman, E., von Davier, A. A., & Hao, J. (2014), *Psychometrics and game-based assessments*. Institute of Play. Retrieved December 27, 2016, from http://www.instituteofplay.org/work/projects/glasslab-research/

Mislevy, R. J., Steinberg, L. S., Almond, R. G., & Lukas, J. F. (2006). Concepts, terminology, and basic models of evidence-centered design. In D. M. Williamson, I. I. Bejar, & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 15–48). Mahwah, NJ: Lawrence Erlbaum.

OECD (2013, March). *PISA 2015 draft collaborative problem solving framework*. Retrieved December 27, 2016, from http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Collaborative%20Problem%20Solving%20Framework%20.pdf

Romero, C., Gonzalez, P., Ventura, S., del Jesus, M. J., & Herrera, F. (2009). Evolutionary algorithms for subgroup discovery in e-learning: A practical application using moodle data. *Expert Systems With Applications*, *39*, 1632–1644.

Shute, V. J., Hansen, E. G., & Almond, R. G. (2008). You can't fatten a hog by weighing it—or can you? Evaluating an assessment for learning system called ACED. *International Journal of Artificial Intelligence and Education*, *18*, 289–316.

Soller, A., & Stevens, R. (2008). Applications of stochastic analyses for collaborative learning and cognitive assessment. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 217–254). Charlotte, NC: Information Age.

VanLehn, K. (2008). Intelligent tutoring systems for continuous, embedded assessment. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 113–138). New York, NY: Lawrence Erlbaum.

von Davier, A. A. (2015, July). Virtual and collaborative assessments: Examples, implications, and challenges for educational measurement. Invited Talk at the Workshop on Machine Learning for Education, International Conference of Machine Learning, Lille, France.

von Davier, A. A., & Halpin, P. F. (2013). *Collaborative problem solving and the assessment of cognitive skills: Psychometric considerations* (Research Report -13-41). Princeton, NJ: Educational Testing Service.

von Davier, A. A., van der Schaar, M., & Baraniuk, R. (2016, December). Workshop on Machine Learning for Education, International Conference of Machine Learning, New York, NY.

von Davier, A. A., Zhu, M., & Kyllonen, P. (Eds.) (2017). *Innovative assessment of collaboration*. New York, NY: Springer Verlag.

Wang, Z., Hao, J., Liu, L., Chen, L., & von Davier, A. A. (2015, July). *Automated classification of collaborative problem solving activities from chat messages*. Paper presented at the 10th annual Interdisciplinary Network for Group Research (INGRoup) conference, Pittsburgh, PA.

Zhang, M., Hao, J., Li, C., & Deane P. (2015, July). *Classification of writing styles using keystroke logs: A hierarchical vectorization approach*. Paper presented at the international meeting of the Psychometric Society, Beijing, China.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web site.

Online Appendix

Data, Log Files, and Data Dependencies