The Gordon Commission
on the Future of Assessment in Education

Toward the Relational Management of Educational Measurement Data

Gregory K. W. K. Chung
University of California, Los Angeles
National Center for Research on Evaluation, Standards, and Student Testing (CRESST)

Author Note

Gregory K. W. K. Chung, National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/UCLA. The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305C080015. The findings and opinions expressed in this paper do not necessarily reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences, or the U.S. Department of Education. I wish to thank Joanne Michiuye of UCLA/CRESST for her help with the preparation or this manuscript. Correspondence concerning this article should be addressed to Gregory K. W. K. Chung, UCLA CSE/CRESST, 10945 Le Conte, 1400C, Box 957150, Los Angeles, California 90095-7150. email: **greg@ucla.edu**.

"Through measurement to knowledge," coined by Kamerlingh Onnes in 1882, succinctly conveys the critical role of measurement in science (Laesecke, 2002). Historically, significant advances in scientific understanding have followed advances in measurement and observation. As the resolving power of an instrument increased, so have gains in the understanding of the phenomena being observed. An example is the microscope, which led to insights, verification, and new research questions and theories over the last 300 years since its invention. The microscope's resolution—the degree of detail that could be distinguished in an image—allowed observation of processes and states that were previously unobservable.

We are now approaching a similar potential in the measurement of students' learning processes using technology-based tasks. Technology-based tasks can be instrumented to record fine-grained observations about what students do in the task as well as capture the context surrounding the behavior. Advances in how such data are conceptualized, in storing and accessing large amounts of data ("big data"), and in the availability of analysis techniques that provide the capability to discover patterns from big data are spurring innovative uses for assessment and instructional purposes. One significant implication of the higher resolving power of technology-based measurement is its use to improve learning via individualized instruction.

Over the last decade there has been renewed interest in the idea of individualized instruction (also called personalized learning). This interest is being driven in part by advances in technology (e.g., ADL, 2009; DOE, 2010; IEEE LTSC, 2006), advances in assessment (e.g., Baker, 1997; Mislevy, in press; NRC, 2001b), and a persistent desire to increase the access, efficiency, and cost-effectiveness of training and education (e.g., Fletcher, Tobias, & Wisher, 2006). Personalization is so significant that it is one of 14 engineering Grand Challenges identified by the National Academy of Engineering (2008).

There is little doubt that as the number of learning applications on digital platforms increases—from transmedia learning apps and games to online assessment—the generation of huge quantities of learning-related data will also increase. If education follows industries that target individuals such as retailing, manufacturing, and healthcare, there will be a widespread interest in how these data could be used to improve student learning and teaching (Bienkowski, Feng, & Means, 2012; Bienkowski, Feng, & Toyama, 2011; DOE, 2010; Manyika et al., 2011). A key issue is how to leverage these data to measure what students understand and can do, to

derive meaningful measures of cognitive and affective processes, and to develop capabilities for precise diagnosis and targeting of instruction.

The remainder of this chapter assumes that there will be a continued increase in the availability of devices with which to observe student behavior and the generation of data associated with students' interaction with those devices. Examples are provided in two broad areas to illustrate how different levels of educational data have been used to understand student learning via a coarse-grained learning analytics approach and via a fine-grained learning process approach. Implications for adaptive systems for teaching and learning are discussed.

Levels of Data Useful for Understanding Student Performance

For the purpose of understanding a student's knowledge and skills, it is helpful to view data at three levels. Each level of data represents different levels of aggregation and can be used to answer different kinds of questions. At the highest level of aggregation is system-level data. For example, the data housed in a student information system (SIS) reflect the indicators important to an institution. In a university, these data would include students' course-taking information, course grades, high school information, and demographic information. These kinds of data allow institutions to ask questions about system-level issues such as student retention rates, graduation rates, and time to degree. Educational measurement often generates individual-level data. Some examples include total score on an achievement test, scores on a performance task, or scores on individual items in a test. In general, this level has been the finest grain-size used in educational measurement. More recently, there has been interest in the use of data at an even finer level of detail and made practical only in technology-based applications (e.g., Romero, Ventura, Pechenizkiy, & Baker, 2011). Transaction-level data reflect a student's interaction with a system where the interaction may be an end in itself (e.g., the action a learner performs as part of gameplay) or a means to an end (e.g., the act of uploading an assignment in a learning management system). In either case, these interactions are increasingly becoming data sources about students' moment-to-moment choices on some task and are beginning to be captured and stored in a format suitable for analyses of student learning.

Examples of Using Data to Individualize Teaching and Learning

In this section two sets of examples are provided to illustrate opposite ends of the data spectrum. The first example is the use of learning analytics in a higher education setting that fuses macro-level data from the institution's SIS with usage data from the institution's learning management system (LMS) to predict individual student success in a course. The second set of examples focus on the derivation of fine-grained measures to understand student learning processes.

Learning Analytics

Learning analytics is the "use of analytic techniques to help target instructional, curricular, and support resources to support the achievement of specific learning goals" (van Barneveld, Arnold, & Campbell, 2012, p. 8). In higher education, learning analytics uses data culled from the SIS, LMS, course grade books and other information sources. Traditionally SIS data have been used to monitor various institutional indicators such as undergraduate retention rates. A recent study that examined the relationship between LMS usage and course outcome found LMS usage to be a good predictor of success when the usage activity was directly related to the learning outcomes of the course (e.g., Macfadyen & Dawson, 2010). The goal of learning analytics is to enable instructors and institutions to tailor the educational experiences of individual students in near real-time. Example applications include predicting student outcomes, creating course dashboards, evaluating curriculum, and identifying students at risk of failing (Bach, 2010). Learning analytics is expected to be implemented in an increasing number of institutions over the next five years (Johnson, Adams, & Cummins, 2012).

One example of learning analytics is the *Signals* project at Purdue University (Campbell & Oblinger, 2007). *Signals* began as an initiative to predict student success (i.e., grades) in a course. *Signals* assumed student success was a function of the student's background (e.g., aptitude, prior experiences) and the student's effort in a course. Student background measures were derived from the SIS, which contained information typical of institutional data (e.g., demographic, SAT®/ACT® scores, high school academic performance, current academic performance, course-taking patterns). Effort was operationalized as the degree to which the student used the university's LMS. The LMS housed course assignments, assessments, and messages. A logistic regression model incorporated these sets of variables to predict course

grades (A/B, or C/D/F). The model predicted student course pass as 80% for freshman students (Campbell, 2007).

One important function of the *Signals* system is to allow an instructor to predict individual student risk. An instructor can run the prediction model for each student in the class. Each student's data are retrieved from the SIS and combined with the current usage data from the LMS. If a student is predicted to be at high risk of not earning an A or B in the course, an intervention can be initiated. The instructor can send students emails and text messages, set up face-to-face meetings, or refer them to academic advisors. Compared to a control group that received no intervention, early and frequent intervention with students identified as at-risk improved their academic performance and help-seeking behavior (Iten, Arnold, & Pistilli, 2008). Feedback that appeared most helpful for student improvement was explicit and focused on improving course outcomes (Tanes, Arnold, King, & Remnet, 2011).

From the perspective of relational data management, the designers of *Signals* asked a key question: Could the information from existing data stores designed for one purpose be used to make point predictions about an individual for the more nuanced purpose of identifying individuals at risk of not being successful? *Signals* exemplifies a practical approach to relational data management to support student learning. Using the institution's existing SIS guarantees the availability of background data as these data are available "for free" and populated regularly from pre-admissions to graduation. Using the institution's existing LMS again guarantees the availability of student engagement data to the extent that students are required to interact with the LMS to be successful in the class. The use of the LMS as a data source is a practical way to incorporate course-specific information into the prediction model to allow tailoring to a specific situation.

Learning Processes

One limitation of general purpose systems designed to support student learning (e.g., an LMS) is that they are designed to host content and not designed to measure learners' interaction with that content. The quality of the measures derived from a general purpose system may be limited to behavior such as frequency of access, time spent on a task, textual data that are entered by students in a discussion board, and student ratings. This is understandable as the systems are designed to provide an infrastructure to support the delivery of a course. However, usage as a

proxy for learning outcomes is less desirable than measures derived from students' direct interaction with a task.

Students' interaction in a task has two dimensions: Outcome measures, which address whether students were able to complete the task, and process measures, which address what students were doing throughout the task. In the first dimension, performance on the task itself is taken as an index of understanding. This has been the traditional approach of online performance assessments (Baker, Chung, & Delacruz, 2008). The second approach is to derive meaningful measures from students' interaction with the system as they attempt to accomplish the objectives of the task. The more the process measures target students' behavior directly relevant to achieving the outcome, the higher the measures' diagnostic value and their potential to predict the outcome.

In the next section, three illustrative examples are drawn from research conducted at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). The examples are illustrative and not intended to be a comprehensive review of the literature. A line of research asked three questions related to online measurement of learning processes: (a) To what extent does students' online behavior relate to their cognitive processing? (b) To what extent can students' online behavior be used to model their problem solving process? and (c) To what extent can students' online behavior be used diagnostically to reveal understandings and misconceptions? The first question was addressed by examining the extent to which measures based on students' online behavior correlated with measures based on their cognitive processes (as measured via their think-alouds). The second question examined user-interface design to extract meaningful data from students' interaction with a simulator. The interface ensured that the measurement reflected a user's intent, which is critical when attempting to model students' problem solving processes. The last question addressed both ideas in a single application: Games for learning. In this case, the extraction of meaningful data is baked into the task itself. Well-designed games are inherently engaging and gameplay behavior is purposeful. When success in the game requires specific domain knowledge and the "game mechanics" require use of that domain knowledge, then the gameplay behavior itself reflects students' understanding. Underlying all three questions is the collection of fine-grained behavioral data and its use to gain insight about students' learning processes.

To what extent does students' online behavior relate to their cognitive processing? In this first example, Chung, de Vries, Cheak, Stevens, and Bewley (2002) investigated the cognitive processes underlying an online problem solving task. Chung et al.'s study was interesting because they used a large sample (89 participants) to examine the relationship between participants' moment-to-moment online behavior and their moment-to-moment cognitive processing. Participants were tasked to help a fictitious friend determine who her "true" parents were from a set of five parents (Stevens & Casillas, 2006). Participants could request various medical tests (e.g., blood type) that would eliminate one or more sets of parents. A library was available that described each test and how to interpret the results. Distracter information in the form of opinions was provided. Participants' online behavior was logged automatically by the web server and participants' were asked to think aloud as they engaged in the task. The think-aloud data were coded and yoked to the web pages participants were viewing; thus, rich behavioral and cognitive processing trace data were available to complement the performance data.

Chung et al. (2002) coded participants' think-aloud data to form categories of cognitive processing (e.g., accurate inferences; inaccurate inferences). Behavioral measures were derived from participants' actions (e.g., examined the test results), and performance measures based on the number of attempts participants took to solve the problem. Chung et al. found significant correlations between students' online behavior and their cognitive processing, particularly with processes that required reasoning (e.g., drawing correct inferences from a series of test results). Participants' online behaviors were quite systematic and purposeful, even when the behavioral data appeared random. For example, what initially appeared as thrashing behavior was explained by participants' think-alouds as them having little idea about what to do in the task. These results in general lend support to the idea that students' online behavior reflects their cognitive processes.

To what extent can students' online behavior be used to model their problem solving process? In the second example, Chung and Baker (2003) investigated whether the "generate-and-test" problem solving strategy could be modeled from participants' interaction with a custom-developed simulator. One underlying problem with well-designed user interfaces is that they may work against the measurement of user intent by making the user interface *too* easy. For

example, measuring whether a student was monitoring time can be difficult because time is typically always visible on a well-designed interface. In their study, Chung and Baker tested a click-through interface where the presence of the information was made obvious, but accessing the information required participants to click through an interface element to reveal the information. This technique allowed tracking of the information participants were actually viewing and the order they were viewing the information.

In the Chung and Baker (2003) study, participants were required to modify a bicycle pump design to meet performance constraints. The information available to the participant was the values of the various pump design variables. The simulation task was instrumented to log the critical events that mapped to components of the generate-and-test model (Jonassen, 2010; Katz, & James, 1998; Newell & Simon, 1972), which involves learners generating an initial solution to a design and then testing that design for adequacy (e.g., design activities, testing the design, inspecting the results, solving the problem). Using sequential analyses, Chung and Baker found that participants' online behavior followed the generate-and-test model. Successful participants tended to access only the information needed to solve the problem rather than view all information for all pump parameters and converged to a solution faster than less successful participants. After incorrect solution attempts, participants tended to examine pump parameters and adjust their pump design, suggesting they were reevaluating the adequacy of their design.

As with the first example, participants' online behaviors were very systematic and mapped to a theoretically-based problem solving process. The current example also illustrated the focus on ensuring as much as possible that the behavior observed is both theoretically guided and reflective of participants' cognitive processing. Requiring users to click through an interface element to view information left little doubt about what participants were viewing because they had to overtly take action to view the information.

To what extent can students' online behavior be used diagnostically to reveal understandings and misconceptions? In the final example, gameplay is being examined as a source of a data for deriving diagnostic information about what students know and whether they possess particular misconceptions in the area of fraction addition. Interestingly, when games are used for learning and assessment purposes, the task can more easily embody the features described by the first two examples compared to simulations or other formats. As a measurement

platform, unique interactions (game mechanics) can be designed to require application of specific domain knowledge thereby serving as a kind of test. An interaction that might be viewed as awkward in a simulation (e.g., the click through interface from the second example) might be more acceptable in a game context if the interaction is part of the game scenario. As a learning platform, designing the game to require use of domain knowledge (e.g., understanding what a denominator is) creates the potential for the game to expose players' knowledge gaps. And when the game mechanic is central to gameplay and success, then the game will evoke cognitive processing directed at beating the game, and evoke sustained and effortful performance from the player (Baker, Chung, & Delacruz, 2012; Shute, Ventura, Bauer, & Zapata-Rivera, 2009; Wainess, Koenig, & Kerr, 2010).

To illustrate these ideas, the final example is from a CRESST-developed game designed to increase students' understanding of the basic fractions concepts of unit and piece size (denominator). The game scenario is to help the character, Patch, move from its initial position to the goal position to free a trapped cat. Patch can only move by following a path specified by the player. The player lays down segments of rope for Patch to follow. The distance Patch travels is determined by the sum of the lengths of the rope segments.

Successful gameplay requires students to first determine the whole unit and then the size of any fractional pieces. The game board is a one-dimensional path (like a number line) or a two-dimensional grid. The grid changes for each game level, which provides students with practice at identifying the unit. Adding rope pieces to the path represents fractional addition and operations are only allowed with like-sized rope pieces. This adding game mechanic corresponds to adding fractions with common denominators. A successful solution results in Patch traveling the appropriate distance to reach the cat, which mathematically is the sum of all rope segment lengths.

The use of gameplay to detect what students do and do not understand is promising for several reasons. In prior research with the game, students have reported the games to be more desirable than traditional classroom instruction and are remarkably engaged in the task. Second, students cannot beat the game without having an understanding of the targeted fraction concepts. The combination of high task engagement and high knowledge requirements creates a testing situation where students exert high effort to beat the level. As suggested by the previous two examples, students' online behavior is systematic and purposeful. However, unlike the second

example where the user interface was intentionally modified to allow for detection of intent (thus imposing a user interface cost to the action), in a game there is more flexibility in how the interface could behave. An action that might be considered cumbersome in a simulation can be perfectly acceptable in a game. And therein lies the potential for the measurement of learning: The creation of game mechanics—the actions in a game that are the essential actions in game play—that also are explicit tests of knowledge (Baker et al., 2012; Plass, Frye, Kinzer, Homer, & Perlin, 2011).

The idea that players' behavior can yield meaningful insight into their understanding has been demonstrated in the fraction game in two ways. First a priori specification of in-game processes, such as correct additions, incorrect addition attempts, and maximum game level reached, all correlated in expected directions with players' prior fractions knowledge and their self-reported math grades (Vendlinski, Chung, Binning, & Buschang, 2011). While these game-based measures provided evidence that the game requires knowledge of fractions concepts, they are not precise about the specific concepts students have difficulty on—math or game.

The second way that players' behaviors can reveal their understanding is in the strategies they use as they attempt to solve a level. Because gameplay can unfold in different ways across different levels, a priori specification of the specific gameplay behavior to examine is difficult. The use of cluster analysis has been particularly helpful in the discovery of gameplay patterns—sets of co-occurring events in a game level. Cluster analysis is a data mining technique useful for extracting learning-related patterns from online behavior (Antonenko, Toy, & Niederhauser, 2012). The cluster analysis has yielded discovery of player errors that verified anecdotal accounts and more importantly, led to the discovery of errors and misconceptions that had not initially been under consideration (Kerr & Chung, 2011; Kerr, Chung, & Iseli, 2011). The two errors that have been observed consistently across studies have been unitizing errors and partitioning errors. Unitizing errors are when students consider a unit as the entire length of a grid by ignoring the intervening unit markers, and partitioning errors, where students count the tick marks that partition the unit and not the number of fractional pieces that make up a unit. These two errors have been found in the literature on fractions understanding to be quite common (NRC, 2001a).

Implications

The idea of increased resolving power based on fine-grained data was presented in the context of relational management of educational data. At one end of the data spectrum was the example of *Signals*, which was used as an example of fusing SIS and LMS data to predict individual student success in a course. At the other end of the data spectrum was the set of examples on the use of fine-grained data on online behavior to infer student learning processes and understanding.

An implication of the availability of such data is the development of adaptive educational systems. General models of adaptation have been established in the literature and are referred to as macroadaptation or microadaptation (e.g., Corno & Snow, 1986; Glaser, 1977; Park & Lee, 2003). Recent work on adaptivity with computer-based technology has focused on the questions of what to adapt and how to adapt (e.g., Shute & Zapata-Rivera, 2012). Macroadaptation refers to the adaptation of the instructional environment to the individual such as a teaching strategy, a prescribed curriculum plan based on students' needs (e.g., an IEP), and allocation of instructional resources and materials to accommodate individual student needs. In microadaptation, the instructional decisions are based on moment-to-moment interactions with a student such as in a one-on-one tutoring situation or in computer-based situations that monitor and respond to a student's ongoing progress on a task. In either adaptation model, the goal is to adapt instruction as precisely as possible to the particular student's need with the goal of maximizing learning outcomes. The role relational educational data has in each of these approaches is described next.

Adaptive Systems to Support Teaching

With respect to supporting teaching, the development and use of data-driven systems designed specifically to support educational improvement in K-12 settings was pioneered by Baker (2000) with the Quality School Portfolio (QSP) system. QSP provided the capability for administrators and teachers to easily import student data at the individual level, including background and demographic information, standardized test scores, benchmark test scores, and whatever other data were available. Graphing, analyses, and a variety of reporting formats were part of the standard QSP services. Unlike *Signals*, QSP was designed to provide more analytical control that would allow administrators and teachers to identify areas in need of improvement specific to the user's district, school, or classroom.

A key feature of QSP was the capability to disaggregate student achievement data, allowing administrators and teachers to conduct data analyses by subgroups. A key finding from a large-scale implementation of QSP was that the large majority of users reported using the system formatively to revise curriculum and instruction. However, the data these decisions were based on were coarse-grained data such as standardized test results and benchmark test results (Heritage, Lee, Chen, & LaTorre, 2005). The general lack of diagnostic information is typically cited as a major shortcoming of the use of achievement tests for instructional use (Heritage & Yeagley, 2005; Herman, Yamashiro, Lefkowitz, & Trusela, 2008).

The reporting of diagnostic information about the particular concepts and misconceptions to the teacher and student may be the area of greatest leverage when using relational management systems. As discussed in the examples, transaction-level data appear to reflect students' underlying cognitive processes and knowledge. When computer-based applications (e.g., games) are designed around learning objectives and purposively instrumented to capture meaningful events, the discovery of strategies and common errors becomes possible using machine learning techniques (Chung & Kerr, 2012). As the number of learning apps and games increase, requiring students to use those apps and games for homework, practice, or other educational purposes may be a compelling way to motivate students to engage educational content while also providing a source of diagnostic information for teachers. Best practices developed by CRESST researchers on how to integrate math games into the curriculum have included identification of common errors exposed by games, what the errors imply about student understanding, and instructional strategies to bolster students' understanding (Vendlinksi & Buschang, 2012).

Adaptive Systems to Support Personalization

One long-term trend educational technology is moving towards is transmedia whereby learners engage in coherent learning experiences across different media platforms. The U.S. Department of Education's Ready To Learn (RTL) program is spurring the development of young children's trans-media apps, which include games integrated with formal learning media and other educational programming. These technologies will be instrumented to capture learners' interactions, and development of measures of learning or proficiency will be based in part on the learner's interaction with the apps (CPB & PBS KIDS, 2011).

A fundamental issue is the technical quality of measures derived from fine-grained data. There has been little empirical research on how to establish the technical quality of such measures and only recently have psychometricians begun to address this issue (see Behrens et al., 2011; Mislevy, in press). Best practices are being developed from various game and simulation researchers (e.g., Chung & Kerr, 2012; Plass et al., 2011; Wainess et al., 2010) on the very first steps of the capture process—what to measure and how to measure it—as this step is often unarticulated and not commonly known, resulting in unusable or uninformative data. A practice recognized as undesirable is to collect "everything" (given the ease with which such data can be collected) and then derive measures post hoc after the data are collected. When developing measures of learning, this approach can be fatal as "everything," unless precisely specified by researchers, will be defined by non-researchers and may not be the indicators related to learning or stored in a format suitable for analysis. Much more research is needed on how to derive meaningful measures from fine-grained data. Chung and Kerr (2012) describe an approach used for a CRESST-developed game that is based on observational research tradition, conceptualizing learners' interactions with the system as a type of behavioral observation. The principal advantage of viewing an interaction as a measurement point is that issues of validity and reliability become central to conceptualizing what to measure, how to measure it, and how the interaction relates to learning processes and outcomes. Precision is gained through an analysis of the interaction with respect to the cognitive demands of the task, which can expose the contextual information surrounding the interaction that will allow interpretation of the interaction.

The use of transaction data is an emerging area of research for educational data mining and learning analytics researchers (e.g., Baker & Yacef, 2009; Romero & Ventura, 2007; Siemens & Long, 2011). For example, progress has been made on developing robust measures of students' online engagement, "gaming" behavior (Baker, Corbett, Roll, & Koedinger, 2008), and off-task behavior (Baker, 2007) in an intelligent tutor system. When part of a formalized assessment design process such as Evidence Centered Design (Mislevy, in press), fine-grained measures derived from students' online interactions can be part of an evidence chain used to generate insight into learners' progression through a task—their ongoing learning, errors, and overall achievement (Behrens, Mislevy, DiCerbo, & Levy, 2011).

The development of robust measures will presumably lead to more effective instructional practices and student learning. Whether diagnostic information is culled from gameplay and reported to teachers to help them decide where to allocate instructional resources, or used in adaptive technology-based systems to "sense" when to provide immediate feedback or execute different instructional branching strategies, the availability of high-quality measures will be critical for any precise targeting of instruction.

In 2003 Gordon and Bridglall lamented the state of affairs where policymakers, researchers, and practitioners depended on aggregated data to make educational decisions. Ten years later, advances in technology have made feasible the large-scale capture, processing, and analysis of students' moment-to-moment interactions with technology-based systems. As education enters the era of big data and transmedia-based learning, the capture of multiple levels of data *for an individual* that spans system-level, individual-level, and transaction-level data will provide new opportunities to understand student learning to a degree of precision rarely possible before, and provide new opportunities to leverage such data to deliver on the promise of personalized instruction.

References

Advanced Distributed Learning (ADL). (2009). Sharable Content Object Reference Model (SCORM), Version 2004 4th Edition. Alexandria, VA: Author.

Antonenko, P. D., Toy, S., & Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. *Educational Technology Research and Development, 60*, 383–398.

Bach, C. (2010). Learning analytics: Targeting instruction, curricula and support services. *Proceedings of the 8th Annual Conference on Education and Information Systems, Technologies and Applications: EISTA 2010*, Orlando, FL.

Baker, E. L. (1997). Model-based performance assessment. *Theory Into Practice, 36*, 247–254.

Baker, E. L. (2000). *Understanding educational quality: Where validity meets technology*. In William Angoff Memorial Lecture Series. Princeton, NJ: Educational Testing Service.

Baker, E. L., Chung, G. K. W. K., & Delacruz, G. C. (2008). Design and validation of technology-based performance assessments. In J. M. Spector, M. D. Merrill, J. J. G. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 595–604). Mahwah, NJ: Erlbaum.

Baker, E. L., Chung, G. K. W. K., & Delacruz, G. C. (2012). The best and future uses of assessment in games. In M. Mayrath, J. Clarke-Midura, & D. H. Robinson (Eds.). *Technology-based assessments for 21st Century skills: Theoretical and practical implications from modern research* (pp. 299–248). Charlotte, NC: Information Age Publishing.

Baker, R. S. J. d. (2007). Modeling and understanding students' off-task behavior in intelligent tutoring systems. *Proceedings of ACM CHI 2007: Computer-Human Interaction*, 1059–1068.

Baker, R. S. J. d., Corbett, A. T., Roll, I., & Koedinger, K. R. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction, 18*, 287–314.

Baker, R. S. J. d., & 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.

Behrens, J. T., Mislevy, R. J., DiCerbo, K. E., & Levy, R. (2011). An evidence centered design for learning and assessment in the digital world. In M. C. Mayrath, J. Clarke-Midura, & D. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research* (pp. 13–54). Charlotte, NC: Information Age.

Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing teaching and learning through educational data mining and learning analytics: An issue brief.* Menlo Park, CA: SRI International.

Bienkowski, M., Feng, M., & Toyama, Y. (2011). *Summary of findings from data mining industry interviews*. Menlo Park, CA: SRI International.

Campbell, J. P. (2007). *Utilizing student data within the course management system to determine undergraduate student academic success: An exploratory study*. Unpublished doctoral dissertation, Purdue University.

Campbell, J. P., & Oblinger, D. (2007). *Academic analytics*. Washington, DC: EDUCAUSE Center for Applied Research.

Chung, G. K. W. K., & Baker, E. L. (2003). An exploratory study to examine the feasibility of measuring problem-solving processes using a click-through interface. *Journal of Technology, Learning, and Assessment, 2*(2). Available from **http://jtla.org**.

Chung, G. K. W. K., de Vries, L. F., Cheak, A. M., Stevens, R. H., & Bewley, W. L. (2002). Cognitive process validation of an online problem solving assessment. *Computers in Human Behavior, 18*, 669–684.

Chung, G. K. W. K. & Kerr, D. S. (2012). *A primer on data logging to support extraction of meaningful information from educational games: An example from Save Patch* (CRESST Report 814). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Corno, L., & Snow, R. E. (1986). Adapting teaching to individual differences among learners. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 605–629). New York: Macmillan Publishing Co.

Corporation for Public Broadcasting (CPB), & PBS KIDS. (2011). *Findings from Ready to Learn: 2005–2010.* Washington, DC: Author.

Fletcher, J. D., Tobias, S., & Wisher, R. A. (2006). Learning anytime, anywhere: Advanced distributed learning and the changing face of education. *Educational Researcher, 36*(2), 96–102.
Glaser, R. (1977). *Adaptive education: Individual diversity and learning*. New York: Holt, Rinehart and Winston.

Gordon, E. W. & Bridglall, B. L. (2003). *Toward a relational data management system for education* (Pedagogical Inquiry and *Praxis*™, No. 4). New York: Institute for Urban and Minority Education, Teachers College, Columbia University & The College Board.

Heritage, M., Lee, J., Chen, E., & LaTorre, D. (2005). *Upgrading America's use of information to improve student performance* (CRESST Report 661). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Heritage, M., & Yeagley, R. (2005). Data use and school improvement: Challenges and prospects. *Yearbook of the National Society for the Study of Education, 104*, 320–339.

Herman, J. L., Yamashiro, K., Lefkowitz, S., & Trusela, L. A. (2008). *Exploring data use and school performance in an urban public school district* (CRESST Report 742). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

IEEE Learning Technology Standards Committee (LTSC) P1484. (2006). *IEEE P1484.12.3., draft 8, extensible markup language (XML) schema definition language binding for learning object metadata*. Retrieved June 1, 2006, from **http://ieeeltsc.org**.

Iten, L., Arnold, K., & Pistilli, M. (2008, March). Mining real-time data to improve student success in a gateway course. *Paper presented at the Eleventh Annual TLT Conference*. West Lafayette, IN: Purdue University.

Johnson, L., Adams, S., & Cummins, M. (2012). *The NMC Horizon Report: 2012 Higher Education Edition*. Austin, TX: The New Media Consortium.

Jonassen, D. H. (2010). *Learning to solve problems: A handbook for designing problem-solving learning environments.* New York: Routledge.

Katz, I. R., & James, C. M. (1998). *Toward assessment of design skill in engineering* (*GRE*® Research Report 97–16). Princeton, NJ: Educational Testing Service.

Kerr, D. & Chung, G. K. W. K. (2011). The mediation effect of in-game performance between prior knowledge and posttest score. In J. Matuga (Eds.), *Proceedings of the IASTED International Conference on Technology for Education (TE 2011)* (pp. 122–128). Anaheim, CA: ACTA Press. doi: 10.2316/P.2011.754-046

Kerr, D., Chung, G. K. W. K., & Iseli, M. R. (2011). *The feasibility of using cluster analysis to examine log data from educational video games* (CRESST Report 790). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Laesecke, A. (2002). Through measurement to knowledge: The inaugural lecture of Heike Kamerlingh Onnes. *Journal of Research of the National Institute of Standards and Technology, 107*, 261–277.

Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. *Computers & Education, 54*, 588–599.

Manyika, J., Chui, J., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity.* Washington, DC: McKinsey Global Institute.

Mislevy, R. J. (in press). *Evidence-centered design for simulation-based assessment.* Military Medicine (Special issue on simulations, H. F. O'Neil, editor.).

National Academy of Engineering (NAE). (2008). *Grand challenges for engineering*. Washington, DC: Author.

National Research Council. (2001a). *Adding it up: Helping children learn mathematics.* J. Kilpatrick, J. Swafford, & B. Findell (Eds.). Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council. (2001b). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Newell, A., & Simon, H.A. (1972). *Human problem solving*. Prentice-Hall, Englewood Cliffs, NJ.

Park, O.-C., & Lee, J. (2003). Adaptive instructional systems. In D. H. Jonassen & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (2nd ed., pp. 651–684). Mahwah, NJ: Erlbaum.

Plass, J. L., Frye, J., Kinzer, C., Homer, B., & Perlin, K. (2011). *Learning mechanics and assessment mechanics for games for learning* (G4LI White Paper 01-2011). New York: NYU/Games for Learning Institute.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 35,* 135–146.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. J. d. (Eds.) (2011). *Handbook of educational data mining*. Boca Raton, FL: CRC Press.

Shute, V. J., Ventura, M., Bauer, M. I., & Zapata-Rivera, D. (2009). Melding the power of serious games and embedded assessment to monitor and foster learning: Flow and grow. In U. Ritterfeld, M. J. Cody, & P. Vorderer (Eds.), *Serious games: Mechanisms and effects* (pp. 295–321). Philadelphia, PA: Routledge.

Shute, V., & Zapata-Rivera, D. (2012). Adaptive educational systems. In P. J. Durlach, & A. M. Lesgold (Eds.), *Adaptive technologies for training and education* (pp. 7–27). New York: Cambridge University Press.

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review, 46*(5).

Stevens, R. H., & Casillas, A. (2006). Artificial neural networks. In D. M. Williamson, I. I. Behar, & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 259–312). Mahwah, NJ: Erlbaum.

Tanes, Z., Arnold, K. E., King, A.S., & Remnet, M. A. (2011). Using Signals for appropriate feedback: Perceptions and practices. *Computers & Education, 57*, 2414–2422.

U.S. Department of Education (DOE). (2010). *Transforming American education: Learning powered by technology.* Washington, DC: Author.

van Barneveld, A., Arnold, K. E., & Campbell, J. P. (2012). *Analytics in higher education: Establishing a common language*. Washington, DC: EDUCAUSE Center for Applied Research.
Vendlinksi, T. P., & Buschang, R. E. (2012, March). *Effectively incorporating video games into math instruction: Results from recent field studies.* Roundtable at the Society for Information Technology and Teacher Education, Austin, TX.

Vendlinski, T., & Stevens, R. (2002). Assessing student problem-solving skills with complex computer-based tasks. *Journal of Technology, Learning, and Assessment, 1* (3). Available from **http://www.jtla.org**.

Vendlinski, T. P., Chung, G. K. W. K., Binning, K. R., & Buschang, R. E. (2011). *Teaching rational number addition using video games: The effects of instructional variation* (CRESST Report 808). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Wainess, R., Koenig, A., & Kerr, D. (2010). *Aligning instruction and assessment with game and simulation design*. Proceedings of the 2010 Interservice/Industry Training, Simulation, and Education Conference, Orlando, FL