Psychometric Considerations for Digital Assessments

David Gibson, *dgibson@simschool.org*
simSchool – Pragmatic Solutions, Westlake Village, CA, USA

Abstract
New psychometric considerations for educational assessments are needed due to two factors: (1) measurement targets that focus on higher order, collaborative and creative thinking skills, and (2) the affordances of digital media learning environments, in particular, the use of interactive game and simulation methods in educational assessments. Interactive applications, many of which utilize social media methods and rely on global storehouses of information, constitute a new genre of assessment performance space. Traces of a learner's problem-solving, expression or communication can now entail highly detailed computer-based documentation of the context, actions, processes and products. Traditional psychometrics, founded in population-based statistics, is proving inadequate to the required tasks of pattern recognition, inference, and visualization of learning and performance. To address the challenge, data-rich, model-driven scientific approaches involving complexity concepts are proposed as part of the theoretical foundation for new psychometric considerations.

Keywords
Digital performance assessment, assessment, psychometrics, computer modeling, simulation-based learning and assessment, learning analytics, complex systems

INTRODUCTION
Today, a person responding to assessment prompts embedded in a digital game or simulation, or anyone working with digital media and tools in an online space, can perform a wide range of actions, can do so continuously over extended periods of time, and can leave behind high-resolution traces of decision-making, intentions, and even emotions (Dede, Nelson, Ketelhut, Clarke, & Bowman, 2004; Gibson, 2011; Quellmalz et al., 2012). These new ways of performing introduce a need for new psychometric considerations for methods of data capture, analysis and display. Digital performance-based data also raise the possibility that learners, teachers, psychometricians and others may be able to assess and evaluate new kinds of assessment targets best exhibited through complex tasks and artifacts, such as critical thinking, collaboration, creativity, communication as well as physical performances that demonstrate these skills. This article explores considerations for the tools and frameworks needed to infer what someone knows and can do from the traces they leave in digital environments.

Psychometrics is the branch of psychology that deals with the design, administration, and interpretation of quantitative tests for the measurement of psychological variables such as intelligence, aptitude, and personality traits (American Heritage Dictionary Online, http://www.ahdictionary.com/). Until recently, the field dealt almost exclusively with the construction and validation of measurement instruments such as questionnaires, tests, and personality assessments; however there is now a need to expand to include highly interactive digital learning and adaptive test experiences. A good psychometric test is "internally consistent, reliable over time, discriminating and of

demonstrated validity in respect of its correlations with other tests, its predictive power and the performance of various criterion groups. It also has good norms." (Kline, 1998 p. 92). The challenge is to preserve these features while reinventing the procedural foundations of psychometrics, which until recently have been exclusively based on population statistics. The new foundation demands time sensitivity, spatial relationships, multiple layers of aggregations at different scales, and a sense of the dynamics of a complex behavior space (Gibson & Jakl, in preparation).

The plan of the article is to ground the discussion in performance assessment, characterize the new performance space of digital media, introduce concepts and approaches from data science and complexity, and discuss why new measurement targets need new analysis tools.

RE-THINKING PERFORMANCE ASSESSMENT

Performance tasks and assessments are defined by (Stecher, 2010):

> A performance task is a structured situation in which stimulus materials and a request for information or action are presented to an individual, who generates a response that can be rated for quality using explicit standards. The standards may apply to the final product or to the process of creating it. A performance assessment is a collection of performance tasks.

The postmodern era of performance assessment (1980 to the present) sounded promising as a way to improve tests and measurements. In Stecher's definition, "*Information* or *action*" and applying standards "to the final *product* or to the *process* of creating it" makes it clear that performance assessment includes an evaluation of knowledge-in-action. Individuals who "*generate* a response" can do more than choose an answer; they can create, communicate and collaborate with others over time. Promisingly, the final products and processes can be "rated for quality." So, performance assessment sounds promising for adding measurement of higher order knowledge and skills to high stakes decisions such as advancing to the next grade or graduating from school.

However, in the two decades before ubiquitous mobile computing and highly interactive digital media, problems with administering a standardized performance assessment for high stakes decisions outweighed the potential benefits. Two problems illustrate why. Complex performances in fields like science demanded access to equipment and lab-like settings as well as processes for capturing the actions and thinking of the test taker on paper or video. Second, storing, organizing and scoring those complex performances involved people not machines, which increased costs while impacting test reliability and fairness. As high-stakes large-scale assessments, the performance assessment systems of those decades were not easy to administer nor cost effective, and were not politically viable (Madaus & O'Dwyer, 1999).

The promising-sounding innovation of large-scale performance assessment thus ground to a halt. Limitations concerning human judgment of paper-based student responses created problems for managing and scaling fair implementations, reliably scoring artifacts, and appropriately reporting results. In Vermont, for example, where statewide portfolios in mathematics and writing were developed during the 1990's, summer scoring sessions involved hundreds of people occupied for a week, judging, moderating

decisions and adjusting scores. While the state's system engaged large numbers of educators in development and implementation and produced positive changes in curriculum, it fell under its own weight a short time after a RAND study indicated that the assessment system had problems with reliability (Koretz, Stecher, Klein, & McCaffrey, 1994). Similar problems befell various aspects of the innovations in Kentucky, Maryland, Washington, and elsewhere (Stecher, 2010).

Without computer-based administration and automated rating, the experience of history seems to tell us, performance assessment is impractical for large-scale summative assessment. But now in the era of ubiquitous computer games and simulations, has the situation changed? Can a computer present a complex situation in a structured way to large numbers of people, capture the open-ended responses generated, and reliably score complex performances? If so, then maybe performance assessment requires re-thinking in the digital age.

A NEW PERFORMANCE SPACE

Digital games and simulations provide learning experiences in a dynamic new performance space with implications for assessment (Mayrath, Clarke-Midura, & Robinson, 2011; Tobias & Fletcher, 2011). The interactions of a game-playing test-taker for example, unfold as performances in *time* and cover a multivariate *space* of possible actions. Neither of these interactions fits the standard model of a multiple-choice questionnaire, which takes place in one slice of time with a highly constrained set of response options. Objective tests are exemplified by items created and curated by experts, and matured in the age of schooling before interactive adaptive computer applications, open access to global resources and a surge in interest in higher order skills such as collaboration and creativity. As Graesser (2012) expresses for nearly anyone who was schooled in the last half of the 20th Century;

> I never played multiparty games that required timely opportunistic communication and negotiation strategies with invisible players. Collaborative problem solving to achieve group goals was not part of our curriculum. I never learned how to manage limited resources and understand trade-offs between factors with an interactive simulation...I was never encouraged or taught how to ask deep questions (why, how, what-if, so what) and to explore novel hypotheses…I had no concept of the value of cross-checking multiple information sources to converge closer to the truth. I never designed or modified a game to make it interesting, challenging, or educational.

When most people take tests, they make choices from limited options such as a, b, c, d, and none or all of the above. They might work math problems and show their results, but not their problem solving process; they might fill out short answers or write essays. Other performances that are more common in school than in large-scale assessments, such as giving a book report, debating, or making a presentation, are often graded subjectively, even when an evaluator uses a scoring rubric that improves reliability. Complex performances are often graded with a pass-fail checkmark, a rubric-based score, or little more than a smile, or a pat on the back. This is the limited extent of the performance space of traditional educational assessment; but now, with interactive digital applications and the Internet, that space has changed.

Similar in some ways to the representation of knowledge called a *problem space* (Fikes & Nilsson, 1971), the new *digital performance space* is a domain-independent representation of the possible ways one can show what one knows and can do. Performance tasks in a digital-media learning environment often have *relatively unconstrained parameters* such as interactions with the mouse, keyboard and screen, storyline choices, open-ended text responses, recorded speech, drawing, use of digital tools (including simulations of real-world tools). In addition, the assessment record can contain historical traces from problem solving decisions, and biometric information such as facial expressions, skin responses and brain states (Brave & Hass, 2003; Park, Baek, & Gibson, 2008; Parunak, Bisson, Brueckner, Matthews, & Sauter, 2006).

However, unlike a problem space, the digital performance space is not primarily a mental model. Instead, it contains both *intangible* (e.g. value, meaning, sensory qualities, and emotions) and *tangible* assets (e.g. media, materials, time and space) that a performer utilizes to communicate a response, bounded by constraints and affordances of the tangible assets, and imagination and intent among the intangibles. Modern approaches to *virtual performance assessment* (Clarke-Midura, Mayrath, & Dede, 2010) are attempting to grapple with a range of new, relatively untapped performance capabilities in the new interactive digital space. Both processes and products are of interest in this new world, but processes in particular are now much more amenable to documentation and analysis than ever before, because events in a digital space can be documented at the micro-second level, producing high-resolution data for visualization and analysis.

In addition to providing complex stimuli and response options delivered to and from the user with high-resolution data, digital assessment tasks can also *dynamically adapt* to the test taker. Assessments can use accurately timed micro-data to determine item difficulty (e.g. how a population performed on a task and at what point during a process a performance might have broken down) to hone the estimation of supportable inferences (e.g. where an individual fits within that population), or compare models of the user's performance with that of experts to diagnose knowledge and skill levels, extend practice, and determine learning and performance trajectories (Quellmalz et al., 2012). Dynamic adaptability implies a degree of *personalization*, measurement of which is relatively unexplored by population-based measurement methods, such as how individual expertise develops differentially within a population that is itself evolving with rapid advances of knowledge and new practices; and how complex measurement snapshots should be integrated over time with the mental models of a test taker. Adaptability entails *interaction and dependency* with the test taker, which raises additional measurement challenges concerning the independence of the stimuli with each other; for example, a reduction in error variance over time, if the adaptive application helps the test taker perform better over time.

Dynamic adaptation in particular raises the possibility of *nonlinear behavior*, because the changed environment, which has responded to the test taker, is now the baseline for the test taker's next move (rather than an independent stimulus with no relationship to the previous stimulus). Information from the test taker's own past has been recycled and is now in the test taker's present. Cycling of information is a condition that can lead to nonlinear behaviors, which then violate the assumption that outputs are a simple linear sum of the inputs. For example, suppose that the test taker likes a certain reward and the reward provides an enhanced opportunity to gain more of the reward, then the test taker's role in the dynamics will nonlinearly accelerate the reward system. This violation

of the conditions for separability and independence calls for a new foundation of psychometrics that is capable of dealing with state-dependent dynamics where conflicting correlations and unpredictable behavior can arise even in a deterministic system following simple rules.

Furthermore, the targets of assessment are changing (Griffin, McGaw, & Care, 2012), from basic skills and declarative knowledge measured separately by independent items, toward the integration of knowledge applied to complex open-ended tasks requiring critical thinking, communication, collaboration, and creativity (Bennett, Persky, Weiss, & Jenkins, 2007; Koch & DeLuca, 2012). This shift toward *higher order thinking* complicates the performance space with constructs that test the boundaries of the sociological (e.g. communication), psychological (e.g. collaboration) and procedural (e.g. creativity and critical thinking) domains of measurement.

These novel information sources, measurement targets and communication formats constitute not only a new performance space for expression but also a larger problem space for educational measurement. With increased complexity in both the stimuli and responses, the measurement options and challenges are numerous, as evidenced by recent research exploring the problems and possibilities of the emerging generation of technology-based assessments (Bennett et al., 2007; Bennett, 2001; Choi, Rupp, Gushta, & Sweet, 2010; Clarke & Dede, 2010; Clarke-Midura et al., 2010; Mayrath, Clarke-Midura, & Robinson, 2012; Quellmalz et al., 2012; Rupp, Gushta, Mislevy, & Shaffer, 2010; Shaffer et al., 2009; Stecher, 2010; USDOE, 2010a).

The measurement challenges represented by the new performance space features (i.e. unconstrained responses, interactive adaptability, nonlinearities, personalization, temporal and spatial features, higher order thinking) demands new psychometric methods (Clarke-Midura, Code, Dede, Mayrath, & Zap, 2012). The field formerly concerned with the construction and validation of measurement instruments such as questionnaires, tests, and personality assessments, must now add the unique issues of game and simulation-based assessments to its purview. But there are no “off the shelf” methods that can be directly applied (Choi et al., 2010); thus there is a need to consider new methods and perhaps a new branch of the science of educational measurement.

DATA-DRIVEN SCIENCE AND COMPLEXITY

Data-driven science and complexity are fundamental to the new considerations, for two reasons. First, while traditional psychometric methods are founded on hypothetico-deductive science (Kline, 1998), those methods cannot discover a new hypothesis or rule that explains particular observations, and the methods have not traditionally dealt with high-resolution time-based data. Data-driven science methods, in contrast, seek to inductively discover rules and patterns. Second, complexity concepts are useful for many questions poorly suited to a linear mechanistic conception provided by traditional science. Two key concepts of the field are *emergence over scale* (new whole system behaviors arise that cannot be predicted by the interactions of the parts) and *self-organization over time* (systems evolve and adapt) (Bateson, 1972; Bechtel, 1993; Holland, 1995; Nicolis, 1989). Complexity methods for example, show how “big data” in complicated networks of relationships represent and explain systems exhibiting dramatic changes over time and distance scales – a criterion of interest when an assessment of a performance is based on high resolution time-sensitive data such as the clicks of a user in a digital space.

These points concerning the new scientific criteria for assessments are evident in the aims and scope of the European Physical Journal (EPJ) Data Science open access journal (Springer, 2012) which articulates the change in scientific worldview that matured in the decade 2000-2010 and is asserted here as critical to an understanding of the direction of psychometrics of the new performance space.

> The 21st century is currently witnessing the establishment of data-driven science as a complementary approach to the traditional hypothesis-driven method. This (r)evolution accompanying the paradigm shift from reductionism to complex systems sciences has already largely transformed the natural sciences and is about to bring the same changes to the techno-socio-economic sciences, viewed broadly.

The search for conceptually new methods to deal with complexity is in alignment with the needs for a science of measurement for the new performance space, including:

- How to extract meaningful data from systems that exhibit increasing complexity
- How to analyze data in a way that allows new insights
- How to generate data that is needed but not yet available
- How to find new empirical laws, or more fundamental theories, concerning how natural or artificial complex systems work

To answer a scientific question, after exploring background and prior knowledge, data-driven science starts with data instead of a hypothesis (Cangelosi & Parisi, 2002). The data-driven science method proceeds by generating, synthesizing or combining large quantities of data from disparate sources, with interchangeable steps of analyzing existing data, or building and running models or simulations that creates data. The core steps of "analyze-model-simulate-generate" in data-driven science contrast with "hypothesize-experiment-analyze" in the traditional scientific method (Steps 3-5 in Table 1). Data-driven science often includes model building, open-ended and discovery-oriented experiments and simulations, data mining, and novel data enrichment via visualization, interpolation, and theory-driven data generation. Experiments, rather than being used as a tightly controlled test of a highly constrained hypotheses, are often used *en mass* to attempt to validate the broader scope and boundaries of an explicitly induced model of a theory in its context, including its simulated or mined data patterns, and its entire set of detailed predictions.

	Traditional Scientific Method	Data-Driven Science Method
Step 1	Ask a Question	Ask a Question
Step 2	Do Background Research	Do Background Research
Step 3	Construct a Hypothesis	Simulate Theory or Search for Patterns
Step 4	Test Hypothesis with Experiments	Generate and Analyze Data
Step 5	Analyze Data and Draw a Conclusion	Conduct Validation Experiments
Step 6	Communicate Results	Communicate Results

Table 1. Comparison between traditional and data-driven science methods

In Step 3 of Table 1, the traditional scientific method posits a hypothesis and submits it to an experiment, while the data-driven method searches for explanatory rules by either creating a computational model that generates new data, or searching for patterns in existing data with algorithmic support from computational agents. Without computational agents, progress in data-driven science would be impossible. The inseparability of computers from the process of scientific discovery characterizes data-driven science and represents an advance over pure mathematics alone as the handmaiden of science (Wolfram, 2002). The advance moves from inert symbols and processes of theoretical mathematics to an interactive, dynamic and practical machine intelligence with superior computational power as a helper in the process of science.

With the new interactivity and computational power has also come the ability of a machine to simulate evolutionary processes (e.g. genetic programming and adaptive algorithms as in (Holland & Reitman, 1978; Koza, 1992)) in order to search complex spaces for rapidly changing rules and conditions. Central to data-driven science methods then, are issues of *machine learning*, how a computer program can discover and check another computer program, adapt its methods to changing circumstances, and *assist humans in finding generalized rules that underlie system structure and behavior*. Koza (1992) for example, holds that (1) virtually all problems in artificial intelligence, machine learning, adaptive systems, and automated learning can be recast as a search for an algorithm, and (2) genetic programming provides a way to successfully conduct the search for an algorithm in the space of all algorithms. The search for and validation of algorithms in data-driven science replaces the role of the experimental result in the traditional scientific method (Step 5 in Table 1).

It is asserted here that the theoretical tools and approaches of big data and complexity will allow the new psychometrics to increase the number and kind of measurement targets that can be considered, just at a time when educational measurement has a renewed interest in new skills needed for success in the 21st Century as well as a new platform for the construction and delivery of assessments.

NEW MEASUREMENT TARGETS

Targets for educational measurement are moving beyond units of knowledge and skill (e.g. remembering a fact from History or Mathematics) toward deeper learning and higher order, more complex, transdisciplinary and socially based knowledge. Three signs of this change in the U.S. are the development of Common Core state standards (NGA & CCSSO, 2010), the Race to the Top Assessment Program (USDOE, 2010b), and private foundations stimulating innovation in digital learning. The Hewlett Foundation's strategic plan for education for example (Hewlett, 2010) defines deeper learning as:

- Mastering core academic content
- Critical thinking and problem solving
- Working collaboratively
- Communicating effectively
- Learning how to learn independently

In addition, the new digital performance space raises additional potential targets, exemplified by the conception of knowledge in the *epistemic frame* of a profession that is present in serious games (Shaffer, 2007). The epistemic frame of a profession is the

collective practitioners' understanding of the skills, knowledge, identity, values, and epistemology of the field of practice. The salience of the new targets arising from games and simulations comes in part from the immersive experience, which provides psychometricians with new kinds of evidence, for example *projective identity* (Gee 2008) that has been defined as a process where embodiment reveals thinking. Note also the added dimensions of *social knowledge* (e.g. professional identity, values and epistemology), which reach beyond the strictly individual focus of traditional psychometric measures of ability, personality and motivation.

Considering the issue of measurement from a foundational and philosophical perspective, the computational epistemologist Paul Thagard points to five sources of knowledge: *explanatory, deductive, conceptual, analogical, and perceptual* (Thagard, 2000). The new psychometrics, bound as it is to the interactive computational engines of digital media-based learning, will need to rise to the challenge of providing scientifically sound educational measures using each of these sources. To do so, it will need to go beyond the current limitations of educational measurements, which focus primarily on deduction (in summative assessments) and explanation (in formative assessments).

Kline (1998) for example, holds that sound measures must have true zeros, and be expressible as ratio scales and asserts "the way forward for the new psychometrics is experimental and biological" (p.180). Kline's objective seems to be to winnow the field of psychometrics to only those measures that fit the factor analysis method of knowledge production, a deductive knowledge source. Two possible responses to Kline are (1) to question whether advancement in all of science operates by only the factor analytic method and (2) to make a case that new measurement techniques and the underlying reality they measure in the learning sciences must meet new criteria for educational assessment targets of interest.

The position taken here is that in order to measure for example, social knowledge or its application, a naturalistic performance environment has to be provided, and a toolset of methods sensitive to the unique affordances of social knowledge-in-performance has to be applied. It is self-evident that the best test of riding a bike would not be a bubble test but an actual performance of riding. Similarly, many of the higher order skills of most interest today need specialized performance opportunities to be elicited and demonstrated. The new digital performance space allows the creation of closer analogs to these real performance spaces than the previous methods available; and thus the need for significant development of the toolset for a new psychometrics capable of helping humans make digital performance assessment judgments.

NEW ANALYTIC TOOLSETS

Space prevents a full discussion of the emerging toolsets, but some broad indications are possible. *Data mining* and *machine learning* are two of the pillars. Within each area are a multitude of techniques and algorithms for finding patterns, discovering production rules, exploring solutions, and constructing models. Exotic phrases such as "nonlinear state space reconstruction," "weakly coupled variables, " and finding the "underlying manifold governing the dynamics" need to be explored and integrated into the analytic frameworks of educational assessment. These new methods suggest that educators need to partner with computer scientists to develop a working knowledge of the field. In addition, new exploratory tools are becoming widely available and should begin to replace the monopoly held by statistical packages. For example, the WEKA project of

Waikato University in New Zealand (http://www.cs.waikato.ac.nz/ml/weka/) and the Euerqa project of Cornell University (http://creativemachines.cornell.edu/eureqa) provide a wealth of readily available tools.

A handful of projects are underway in the U.S. exploring the new psychometrics of digital learning. A recent book (Mayrath et al., 2012) contains reports from some of them. For example Debbie Denise Reese, in the Selene project (http://selene.cet.edu/), has discovered the performance signature of the "ah-ha" moment when users learn how to utilize the game mechanic to cause planets to form. Jody Clarke-Midura of Harvard and I are working together to analyze performance data concerning scientific reasoning by middle school students who played a game requiring them to collect and utilize evidence to build a scientific argument. A pilot project by the Educational Testing Service is creating a suite of tools to diagnose and identify the learning needs of English Language learners, then instruct them and determine when they should exit special services. Integral to the suite are new psychometric measures of language performance that take advantage of the affordances of the digital performance space; new conceptions of tasks, test prompts, and test taker responses are evolving. As these projects mature, the new science of psychometrics will also evolve.

CONCLUSION

In an interactive digital-based assessment, traces of a learner's problem-solving attempts, self-expressions and social communications can now entail highly detailed computer-based documentation of the context, actions, processes and products. New psychometrics are needed to address the challenges of finding patterns and making inferences from this data. Data-rich, model-driven scientific approaches involving complexity concepts are proposed as a theoretical foundation for new psychometric considerations that include time sensitivity, spatial relationships, multiple layers of aggregations at different scales, and a sense of the dynamics of a complex behavior space.

References

Bateson, G. (1972). *Steps to an ecology of mind: Collected essays in anthropology, psychiatry, evolution, and epistemology* (p. 545). San Francisco: Chandler Pub. Co.

Bechtel, W. (1993). *Discovering complexity: decomposition and localization as strategies in scientific research*. Princeton, NJ: Princeton University Press.

Bennett, R. (2001). How the Internet Will Help Large-Scale Assessment Reinvent Itself. *Education Policy Analysis Archives*, *9*(5), 1–23.

Bennett, R., Persky, H., Weiss, A., & Jenkins, F. (2007). *Problem solving in technology-rich environments: A report from the NAEP technology-based assessment project. The Nation's Report Card* (p. 180). Retrieved from http://nces.ed.gov/nationsreportcard/pdf/studies/2007466.pdf

Brave, S., & Hass, C. (2003). Emotion in human-computer interaction. In J. Jacko & . Sears (Eds.), *The Human-computer Interaction Handbook: Fundamentals, Evolving Technologies and Emergine Applications*. (pp. 81–96). Mahwah, N.J.: Lawrence Erlbaum Associates, Inc.

Cangelosi, A., & Parisi, D. (2002). Computer Simulation : A New Scientific Approach to the Study of Language Evolution. In A. Cangelosi & D. Parisi (Eds.), *Simulating the*

Evolution of Language (pp. 3–28). London: Springer. doi:http://dx.doi.org/10.1007/978-1-4471-0663-0_1

Choi, Y., Rupp, A., Gushta, M., & Sweet, S. (2010). Modeling learning trajectories with epistemic network analysis: An investigation of a novel analytic method for learning progressions in epistemic games. In *National Council on Measurement in Education* (pp. 1–39).

Clarke, J., & Dede, C. (2010). Assessment, technology, and change. *Journal of Research in Teacher Education*, *42*(3).

Clarke-Midura, J., Code, J., Dede, C., Mayrath, M., & Zap, N. (2012). Thinking outside the bubble: Virtual performance assessments for measuring complex learning. In *Technology-based Assessments for 21st Century Skills: Theoretical and Practical Implications from Modern Research*. (pp. 125–148). Charlotte, NC: Information Age Publishers.

Clarke-Midura, J., Mayrath, M., & Dede, C. (2010). Measuring Inquiry : New Methods , Promises & Challenges. *Library*, *2*, 89–92.

Dede, C., Nelson, B., Ketelhut, D., Clarke, J., & Bowman, C. (2004). Design-based research strategies for studying situated learning in a multi-user virtual environment. Cambridge, MA: Harvard University. Retrieved from http://muve.gse.harvard.edu/muvees2003/

Fikes, R., & Nilsson, N. (1971). STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, *2*, 189–203. doi:http://dx.doi.org/10.1016/0004-3702(71)90010-5

Gibson, D. (2011). Modeling Emotions in Simulated Learning. In *Standards in Emotion Modeling*. Lorentz Center International Center for workshops in the Sciences.

Griffin, P., McGaw, B., & Care, E. (2012). *Assessment and teaching of 21st Century skills*. (P. Griffin, B. McGaw, & E. Care, Eds.). Dordrecht: Springer.

Hewlett. (2010). Education program strategic plan.

Holland, J. (1995). *Hidden order: How adaptation builds complexity*. *Helix Books*. Cambridge, MA: Perseus Books.

Holland, J., & Reitman, J. (1978). Cognitive systems based on adaptive algorithms. In D. Waterman & F. Hayes-Roth (Eds.), *Pattern-directed Inference Systems*. New York: Academic Press.

Kline, P. (1998). *The new psychometrics: Science, psychology, and measurement*. London: Routledge.

Koch, M. J., & DeLuca, C. (2012). Rethinking validation in complex high-stakes assessment contexts. *Assessment in Education: Principles, Policy & Practice,*, 1–18.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice*, *13*(3), 5–10. doi:http://dx.doi.org/10.1111/j.1745-3992.1994.tb00443.x

Koza, J. R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*. *Springer* (p. 819). MIT Press. Retrieved from http://books.google.com/books?id=Bhtxo60BV0EC&pgis=1

Madaus, G., & O'Dwyer, L. (1999). Short history of performance assessment: Lessons learned. *Phi Delta Kappan*, *80*(9), 688–695.

Mayrath, M., Clarke-Midura, J., & Robinson, D. (2011). *Technology-Based Assessments for 21st Century Skills: Theoretical and Practical Implications from Modern Research* (p. 386). IAP.

Mayrath, M., Clarke-Midura, J., & Robinson, D. (2012). Introduction to Technology-based assessments for 21st Centry skills. In M. Mayrath, J. Clarke-Midura, D. Robinson, & G. Schraw (Eds.), *Technology-based Assessments for 21st Century Skills: Theoretical and Practical Implications from Modern Research*. Charlotte, NC: Information Age Publishers.

NGA, & CCSSO. (2010). Common core state standards. *National Governors Association Center for Best Practices, Council of Chief State School Officers*. Washington D.C.: National Governors Association Center for Best Practices, Council of Chief State School Officers. Retrieved from http://www.corestandards.org/the-standards

Nicolis, G. (1989). *Exploring complexity: an introduction*. *New York,NY: WH Freeman*.

Park, H., Baek, Y., & Gibson, D. (2008). Design and development of an adaptive mobile learning management system based on student learning styles. In J. Lumsden (Ed.), *Handbook of Research on User Interface Design and Evaluation for Mobile Technology* (Vol. 1, p. 1240). Hershey, PA: Information Science Reference.

Parunak, H., Bisson, R., Brueckner, S., Matthews, R., & Sauter, J. (2006). A model of emotions for situated agents. In *Autonomous Agents and Multi-Agent Systems*. Hakodate, Hokkaido, Japan.

Quellmalz, E., Timms, M., Buckley, B., Davenport, J., Loveland, M., & Silberglitt, M. (2012). 21st century dynamic assessment. In M. Mayrath, J. Clarke-Midura, & D. Robinson (Eds.), *Technology-based Assessments for 21st Century Skills: Theoretical and Practical Implications from Modern Research* (pp. 55–90). Charlotte, NC: Information Age Publishers.

Rupp, A., Gushta, M., Mislevy, R., & Shaffer, D. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *Journal of Technology, Learning, and Assessment*, *8*(4), 1–45.

Shaffer, D. (2007). *How computer games help children learn*. New York: Palgrave Macmillan.

Shaffer, D., Hatfield, D., Svarovsky, G., Nash, P., Nulty, A., Bagley, E., … Mislevy, R. (2009). Epistemic Network Analysis: A Prototype for 21st-Century Assessment of Learning. *International Journal of Learning and Media*, *1*(2), 33–53. Retrieved from http://www.mitpressjournals.org/doi/abs/10.1162/ijlm.2009.0013

Springer. (2012). EPJ Data Science. *Springer Open Access Journals*. Retrieved from http://www.springer.com/computer/information+systems+and+applications/journal/13688?cm_mmc=sgw-_-ps-_-journal-_-13688

Stecher, B. (2010). *Performance assessment in an era of standards-based educational accountability*. *Assessment* (p. 46). Standford, CA.

Thagard, P. (2000). *Coherence in thought and action*. Cambridge, MA: MIT Press.

Tobias, S., & Fletcher, J. D. (2011). *Computer Games and Instruction*. Information Age Publishing.

USDOE. (2010a). *Learning Powered by Technology: Transforming American Education*. *Educational Technology* (p. 88).

USDOE. (2010b). Race to the top assesment program. Washington D.C.: U.S. Department of Education. Retrieved from http://www2.ed.gov/programs/racetothetop-assessment/index.html

Wolfram, S. (2002). *A new kind of science*. Champaign, IL: Wolfram Media.