

SOME PSYCHOMETRIC AND DESIGN IMPLICATIONS OF GAME-BASED LEARNING ANALYTICS

David Gibson¹ and Jody Clarke-Midura²

¹*Curtin University, Perth, WA Australia*

²*The Education Arcade, MIT, Cambridge, MA USA*

ABSTRACT

The rise of digital game and simulation-based learning applications has led to new approaches in educational measurement that take account of patterns in time, high resolution paths of action, and clusters of virtual performance artifacts. The new approaches, which depart from traditional statistical analyses, include data mining, machine learning, and symbolic regression. This article briefly describes the context, methods and broad findings from two game-based analyses and describes key explanatory constructs use to make claims about the users, as well as the implications for design of digital game-based learning and assessment applications.

KEYWORDS

Virtual performance assessment, learning analytics, game-based psychometrics, data mining, machine learning, simulation

1. INTRODUCTION

The growth of digital game and simulation-based learning and assessment applications has given rise to new considerations about how to make sense of what a user knows and can do based on an analysis of interaction log files. The log files are typically a time-stamped record of all the actions taken by the user in the digital space, so they often provide a high-resolution view of the user's performance over time. The data files can become quite large in comparison to typical educational measurements. For example, it is not uncommon to have thousands of records for a single user's thirty minutes of virtual performance interaction, compared with a dozens or perhaps a hundred responses from a thirty-minute multiple-choice "test." Several recently edited books have begun to bring together findings from researchers who are grappling with the issues of time, sequence, action relevancy, big-data pattern recognition, grain size and resolution, overlapping patterns, levels of meaning and other intriguing challenges (Ifenthaler, Eseryel, & Ge, 2012; Mayrath, Clarke-Midura, & Robinson, 2011; Tobias & Fletcher, 2011). New reports from exploratory research can offer additional insight and help lead to ideas that may prove useful for a synthesis of methods that are emerging to deal with the data from interactive digital learning applications.

This article briefly describes the context, methods and broad findings from two game-based analyses and provides an abstract of key explanatory constructs utilized to make claims about the users, as well as the implications for the design and measurement of digital game-based learning and assessment applications.

2. LOG FILES AND LEARNING ANALYTICS

The data for the analyses described here comes from two virtual performance assessments (VPAs) developed by the Virtual Assessment Project at the Harvard Graduate School of Education. The VPAs assessed middle school students' abilities to design a scientific investigations and construct a causal explanation (Clarke-Midura, Dede, & Norton, 2011). The assessments were designed in the Unity game engine (<http://unity3d.com/>) and have the look and feel of a videogame (Figure 1).



Figure 1. Screen shots of the Virtual Performance Assessments (VPA)

The assessments start out with one of two problems that students must solve: Why is there a frog with six legs? What is causing a population of bees to die? Students walk around the virtual environment and visit farms, talk to farmers, collect data, test the data in the lab, and conduct research until they have gathered enough evidence to support a claim that allows them to identify the causal factor.

Every action by every user (e.g. opening a page, saving a note) was time-stamped as an event. The data from pilots of the assessments used in the analysis reported here consisted of 1987 users (423616 event records) in the frog assessment and 1958 users (396863 event records) in the bee assessment. The data examined in this analysis included the raw event data (up to when they make their final claim about the problem) and the scored constructs: designing a causal explanation and designing a scientific investigation. The scored data was scored using a rubric generated by researchers. The scored data was stored in a file that contained demographic information about students (age, gender, class, teacher) as well as their starting prediction for the cause of the problem.

Designing causal explanation is defined as the student's ability to support their claim or conclusion with evidence. The measure of students' ability to design a causal explanation (DCE) was operationalized through assigning points based on whether the evidence they provided supported the claim they made. Students were first asked to identify data that was evidence based on what they collected in their backpack and the tests they conducted. They were then allowed to choose from all possible data in the virtual environment, to give students who may not have collected all the necessary data a chance to support their claim with evidence. Then the student indicated for each piece of data whether or not it was evidence for their claim/conclusion, as well as identifying which farm was causing the problem. Most of the evidence and the final conclusion or claim were scored on a scale of (3, 2, 1, or 0 points). A backpack of objects populated by the student contained up to 5 pieces of data, each worth (3, 2, 1, or 0 points). Overall, DCE is scaled between 0 and 24.

Designing a scientific investigation (DSI) is defined as the student's ability to carry out an investigation to gather evidence to support their claim. The measure of students' DSI ability was operationalized through assigning points based on whether they conducted tests in the labs, used controls, conducted multiple samples, and reviewed informational research on the causal factors. These processes were scored dichotomously, if the students performed the action they were awarded a point. If they did not, they received a 0. Overall, DSI is scaled between 0 and 24. This construct was an attempt by the researchers to turn student investigative processes captured in the log data into products.

The raw event data contained the time-stamped actions students took from the moment they logged in until they were ready to make their final claim, which is called the "event file." The event file had multiple records per user based on the number of events triggered by the user during a single testing session, and contained fields including the time-stamp of each event, a code for the zone, action and object of each event (i.e. where in the application, using what interaction method, and on what objects associated with each event), and the results of in-world interactions that produced a result. For example, if the user conducted a blood test, there might be five testing results showing which tests were performed on frogs and the result of each test. Similar data were available for the bee assessment.

The purpose of the analysis was to search for patterns of action that might relate to the performance of the user correlated with the student's final claim. Could the log and score data tell us about the user's performance? Ultimately can performance in a virtual performance assessment replace performance on a test? Additional questions included:

- Is there a relationship between overall duration and score level?
- Were there performance differences that differed by gender, age, and grade?
- Was there a relationship between someone's prediction at the beginning of the assessment and their claim at the end?
- Did students have different patterns of behavior and resource utilization that were predictive of their claim?
- Were patterns of behavior and resource utilization related to their predictions?

3. TOOLS AND METHODS

Software tools used in the analyses included Excel, Weka, Eureqa and GraphViz (Table 1). Excel pivot tables were used to explore counts and cross-tab relationships among variables. From the pivot tables, subsets of data were exported to Weka or Eureqa depending on whether the goal of the exploration was data mining with cluster methods or symbolic regression. Weka was used to visually inspect data relationships, classify datasets, and discover clusters and association rules. Eureqa was used to conduct symbolic regression searches for mathematical expressions that could best capture the dynamic relationships among the variables under study. GraphViz was used to create network digraphs of the association rules found for subgroups and the population as a whole. At the end of this article is a reflection on the strengths and weakness of each of these tools and their relationship to the overall analysis.

Table 1. Tools Used in Game-Based Analysis

Software	URL	Uses
Excel	http://office.microsoft.com/en-us/excel/	Counts, Pivot Tables
Weka	http://www.cs.waikato.ac.nz/ml/weka/	Data mining with machine learning algorithms
Eureqa	http://creativemachines.cornell.edu/eureqa	Symbolic Regression
GraphViz	http://www.graphviz.org/	Network graphs

In this section, examples of the various methods employed are offered as brief introductions to the approaches and the primary purpose for selecting each one. The goal here is to set the stage for commenting on the psychometric implications by giving an overview of the resulting information obtained with each method and its potential relationship to understanding what a user knows and can do being inferred from the log file of a virtual performance assessment.

3.1 Symbolic Regression

To answer the question about whether duration of performance was related to score, a traditional approach might be to seek a correlation and explain the shape of the data from the point of view of the population as a whole. In contrast, the symbolic regression method using Eureqa (Schmidt & Lipson, 2009) was selected to attempt to obtain a detailed mathematical expression that would be predictive for any individual score given the user's duration in the digital assessment (or vice versa). Correlation in this case is used as a criterion for the fit of the discovered equation. To preprocess the data, information on duration and total score was smoothed and normalized, and outliers were removed (Figure 2). Note the cyclic data relationship as the total score increases from left to right; this cyclic aspect in the data is due to the nature of the time-stamp, which uses modulo math (e.g. the 13th hour resets the hour clock to 1, the 61st second resets the seconds clock to 1, the 61st minute is an increase of 1 hour, with a resulting zeroing out of the variable and thus a cycle). The zero point on the horizontal is the comparable means of the two variables after normalization. A search was performed until equations converged, with Eureqa's default settings for error (squared error) and simple arithmetic expressions (basic operations plus trigonometric building blocks). The selected solution (Eq 1) on the Pareto Curve had $r^2 = .72$ and correlation of .85 (Figure 3 lower right hand corner). The Pareto Curve represents the trade-off in efficiency between error and complexity: the less complex the mathematical expression, the higher the error and vice versa. This example of the use of Eureqa provides evidence for the finding that complex nonlinear relationships can be discovered via symbolic regression.

Eq 1: $total_score = 49.72 * duration / (0.01552 + duration)$

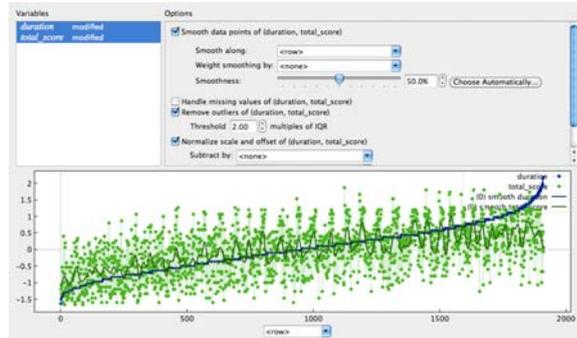


Figure 2. Pre-processing Visualization in Eureka Showing Smoothed and Normalized Data for Duration and Total Score.

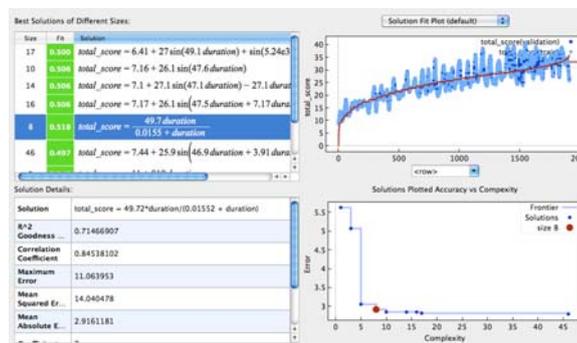


Figure 3. Relationship of Total Score to Duration: Solution on the Pareto Curve

3.2 Counts

To characterize the relationship of prediction to claim, the next example shows the use of the Pivot Table in Excel to display a count of the unique occurrences of users (by student ID) in a cross-tab matrix of prediction versus claim (Table 2). The Pivot Table automates selected mathematical and string operations on variables and arranges the results in a matrix that allows quick exploration of the data.

Table 2. Unique student ID matrix of counts for prediction and claims in A2

Count of student_id	claim_id						
prediction	aliens	mutation	parasites	pesticides	pollution	Total	
.		7	12	14	10	43	
aliens	21	7	11	21	12	72	
dunno	18	50	270	310	150	798	
mutation	5	126	168	209	103	611	
parasites		5	66	40	9	120	
pesticides		1	35	95	30	161	
pollution	1	3	26	65	85	180	
Total	45	199	588	754	399	1985	

Note that 72 students predicted “aliens” as the cause of the unusual frogs (the total of the aliens row), but only 21 of those offered “aliens” as their claim (the intersection of the row with the aliens claim column). The count table provides the basis for *empirical probabilities* based on the ratios of students located at the intersection of prediction and claim choices. In a similar population of middle school students, we would expect that 754/1985 or 38% will likely claim “pesticides.” We can also see that only 8% predicted that result

at the beginning of the assessment, so a significant portion of the test takers arrived at this conclusion (and all other conclusions) after interactions in the virtual assessment. This indicates that the assessment might be educative and that user actions might give clues to a user's thought processes.

3.3 Rule Discovery with Machine Learning

As an example of rule discovery with the aid of machine learning algorithms followed by network analysis, a search of the frog assessment data found the ten best association rules using the "Apriori" algorithm in WEKA (Witten & Frank, 2005). This algorithm seeks to find patterns in an exhaustive search of the data, which can then be used to produce a rule network for the population. Confidence levels are interpreted as the probability of finding the noted association rule within this population.

3.4 Network Analysis

A digraph was then created with GraphViz (<http://www.graphviz.org/>) based on the ten best association rules, which depicts the network of relationships in the data (Figure 4). A digraph is a "directed graph" where the edge from one node to another has a directional meaning – as in causality or implication. An association rule network has directionality if there is not a second rule pointing back from a second node to the first. For example rule one points from research_3 to research_1 with high confidence, but there is no rule pointing from research_1 back to research_3 within the top ten rules, so the digraph captures the one-way relationship and implies that there is more than an association between these two nodes in this particular direction. Such would be the case, for example, if many other users scored a "one" on research_1 but then never interacted with research_3. Experts familiar with the structure of the virtual performance assessment have to validate whether the one-way causal or implicative relationship is appropriate for the structure of the virtual space.

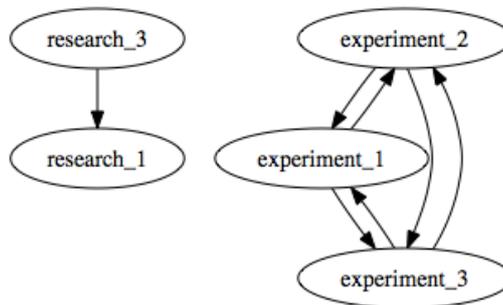


Figure 4. Digraph of Best Association Rules for the Total Population

The rule and network analysis of the frog assessment led to the observation that subgroups that did not have a structure of scientific investigation similar to Figure 4 were more likely to have missed important evidence and reached a weaker conclusion. If this information were used during the assessment to re-direct students to important evidence, then the digital experience would potentially be formative for developing their abilities to design scientific investigations.

3.5 Cluster Analysis

As a final example of analysis methods, we used cluster analysis to find out if there was a relationship between salient moves and clusters formed from all other data in the record. Salient moves, which had been determined by an expert panel, were identified as part of the conceptual framework of the assessment. Each salient move counted as "1" in this analysis, which searched for the number of such moves in relationship to claim and the user's closest cluster using all data in the record. Closeness was determined using the "Expectation Maximization" algorithm (Dempster, Laird, & Rubin, 1977). Clusters mapped closely to claims, but were more complex, because they were formed from all available data. For example, students who shared similar search and resource utilization strategies might be clustered together, even though they reached different conclusions about the data and made different claims (Figure 4).

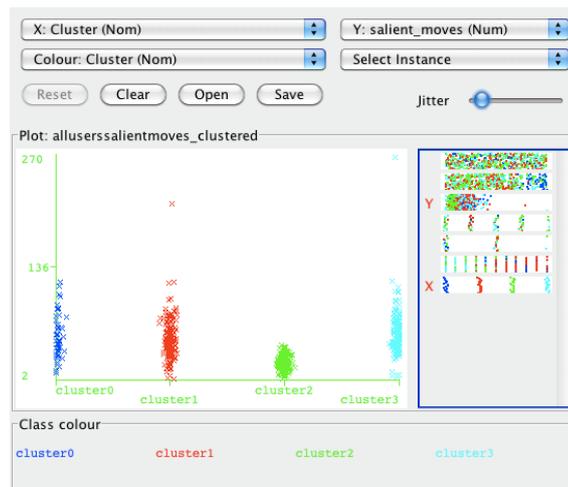


Figure 5. Number of Salient Moves Vs Cluster Membership

It is clear from Figure 5 that cluster 2 used far fewer salient strategies than others, indicating that above a certain number of salient moves, we can predict which cluster a student is NOT a member of, narrow down to the remaining groups and use group probability distributions to estimate other aspects of the student's performance such as total score and final claim.

The above examples of symbolic regression, counts establishing empirical prior probabilities, visualizations, rule discovery, network structure, and cluster analysis methods illustrate some of the new array of tools used in game-based data analysis. In the following sections, a comparison of the methods is presented followed by a summary of main findings and thoughts concerning the design of game-based data collection for educational measurement.

4. EXPLANATORY CONSTRUCTS AND REFLECTIONS

This section briefly highlights the main strengths and weaknesses of the analysis approaches when applied to log data collected from a virtual performance assessment. Overall, the strengths of all the above methods come from their use in *model building* contrasted with *hypothesis testing* and traditional statistical testing. The methods are useful when the questions about data are open-ended and ill-structured; more along the lines of “what have we got here?” than “to what extent is there an impact?”

Symbolic regression (Schmidt & Lipson, 2009) discovers sets of mathematical expressions that capture the dynamics and structures in data, but leaves the decision to the user concerning which expression is best for a particular purpose. The expressions are arrayed from most simple with highest error to most complex with least error, and there is a danger that if an analyst chooses complexity and low error, and performs no other tests or explorations, then an “over-fit” expression will be the result. To ameliorate that threat, cross-validation methods are used; random subsets of the data are used to train as well as test the fitness of the solution. The method is most naturally used with continuous quantitative data, but additional methods can be applied to deal with qualitative data. Groups of such expressions can lead to rule sets, network representations and analysis.

The discovery of association rules among qualitative data can also lead to network representations and analysis (Han, Cheng, Xin, & Yan, 2007). Algorithms in data mining toolsets perform exhaustive searches and optimization routines that result in a descriptive and associative rule set (compared to the mathematical rule set of the symbolic regression method). Such an associative rule set, when considered with the confidence of the rule, can elucidate the hierarchal as well as temporal structure (Campanharo, Siner, Malmgren, Ramos, & Amaral, 2011) of the relationships in a virtual performance assessment created by the paths of multiple users traversing the space and utilizing resources. A limitation of this method is that it is used solely with qualitative data, so continuous data would need to be quantized (Miles & Huberman, 1994) before applying the methods. Log data of a qualitative nature does not need to be coded into numeric bins, as is the case when using SPSS methods.

Visualization methods have been traditionally thought of as the display of findings, so it needs to be emphasized here that visual exploration is itself a form of inquiry as well as demonstration. See for example (Wolfram, 2002) for an example of exhaustive visualization as demonstration. The strengths of the method include the fact that humans have highly evolved visual sense, which facilitates insights from multiple representations and expands understanding of relationships. Thus, WEKA for example, displays visualizations early in the process of data exploration rather than as the last step after analysis. The main weakness is that visualization alone is not enough specific information to convince one of a relationship; so multiple methods need to be combined to tell a complete story of the data.

These strengths and weaknesses are related to explanatory constructs suggested by the VPA data and shared here to stimulate thinking and discussion. In both the frog and bee assessments, a count of unique student ID's on a table of prediction vs claims produced a basis for what might be called the *ecological rationality* (Gigerenzer, Todd, & ABC, 1999) of the performance space, a foundation for computing the joint probability of variables, for example as the *a priori* probabilities in a Bayesian analysis. Empirical probabilities computed from the counts provided a basis for making inferences about the cognitive states of the population viz the affordances of the space and in relationship to the world outside of the performance space. For example, a count of the change from prediction to claim options documented a shift in opinion, implying that the structure of the choices as well as the associated action patterns of the populations making those choices provided evidence of the thought processes that accompanied those decisions. The goal of analysis is then to reconstruct the most likely explanations of action patterns given the ecological rationality of the population.

A second observation is that saliency of a particular action is not a property of the action alone, but has to be paired with an object, the *action-object pair*, as well as a *context of the action*, which leads to the idea of larger action phrases or *motifs*. For example, "opening" any page is an action, but "opening the pesticides page" is a specific action-object pair with more meaning. Furthermore opening that page near the end of the assessment when making the decision about which claim to assert, further contextualizes the meaning of that action. The evidence from the two virtual performance assessment analyses suggests the possibility that the larger the unit of appraisal or motif, the easier it is to discover the variations in the action patterns of users; and the more the context is understood, the more that saliency can be associated with some particular intention or goal.

5. PSYCHOMETRIC IMPLICATIONS FROM THE ANALYSES

Psychometrics involves two major tasks: the construction of instruments and procedures for measurement and the development and refinement of theoretical approaches to measurement. With the advent of game and simulation-based applications for learning, the instruments and procedures for measuring learning and performance are shifting away from point-in-time (e.g. means taken on a slice of time) to patterns-over-time methods (e. g. trajectories evolving during some period of time). This moves the discussion around assessment from numbers to the structure of reasoning (Mislevy et al, 2012). The study presented here illustrates some of the implications of the new methods for theory and procedures needed to imply and estimate what someone knows and can do from game-based actions. The implications fall into several types: nonlinear relationships, rule networks, Bayesian probabilities, and semantic structure of actions.

With log data from an educational game or simulation, complex nonlinear relationships can be discovered via symbolic regression and mathematically expressed with a good degree of precision. For example, in both the frog and bee assessments, a relationship was discovered between duration and score and could be expressed with precision.

Differing performance strategies used by subgroups have a discoverable well-defined meaning and expression in terms of association rules and network structure that can be validated with performance outcomes and scoring. The relationships are complex, overlapping and nonlinear. However, if there is no constraint on utilizing resources in the virtual performance assessment space, then there will be considerable overlap of patterns by all users (everyone uses everything), making the discernment of action patterns more difficult.

Rule networks can be discovered that are useful for making automatic inferences within the constraints of the rule's confidence level. The rules in the VPA for example could classify that the student belongs in or is excluded from a particular performance group, or if the student was already known to be in a performance

group, then when time or action sequences are added to the analysis, scores can be inferred. Tuning up rule mappings requires people who are knowledgeable of the performance space affordances to make adjustments for causal and concurrent influences. Once tuned up, the rule network can help define a multileveled perceptron that can automatically categorize inputs within the constraints of cross-validation results.

Prior probabilities for Bayesian scoring and other automated analyses can be based on the prior performances of cross-validation groups. In the VPAs for example, predictions and claims data from the tested population provided a number of prior probabilities.

Patterns of action-object use (i.e. semantic structure) have predictive value, and we suspect that larger and larger phrases and sequences of action-objects will increase their value by enlarging the salience of the actions. In the VPAs the digraphs and association rules for action-objects provide details for differences in strategy patterns, subgroup membership, and performance level. For example, the pesticides group spent more time than others inspecting the red frog, inspecting the green water, and discarding green water. The pollution group talked to the red farmer more than others. The parasites group talked to more to the scientist and the farmer from the yellow farm than others. These sorts of differences in action-object sequences can be used to classify a user during an assessment.

6. IMPLICATIONS FOR DESIGN OF VIRTUAL PERFORMANCE ASSESSMENTS

The following suggestions based on the analysis and findings reported here are intended to heighten the potential variation among test takers of a virtual performance assessment so that differences in strategies and resulting actions will stand out during analyses.

Designers should plan for larger units of appraisal than the single record event with a time-stemp. If possible, build these recognized units into the application's data collection mechanisms as second order appraisals. The automated recognitions can contain some noise and can also be constrained by windows of time within which all the constituent action-objects must appear, including "the in-sequence appearance" of action-objects when necessary. Related to this observation, there needs to be a method for identifying action sequences that are unique to specific searches and solutions in the virtual performance space.

Time measures should, in addition to time-stamping events, document the event duration (e.g. the start, duration, and ending as a unit) of salient action sequences and use a non-cyclical amount of time, to avoid introducing cyclical artifacts caused by the modulo mathematics of clocks.

In visualizations, utilize an assessment frame-based reason to place sections of action-object pairs closer to each other. For example, if the variables were organized into action-object groupings related to the conceptual structure of the assessment, then time-based visualizations would reveal new patterns and insights.

To avoid the problem of everyone displaying a similar "use everything" strategy, designers should consider utilizing "anti-scoring" penalties that would further restrict the range of scores to better align with highly valued action-object sequences OR have clearly defined outcome subscales that align with scores. Resource utilization behavior would change if there was a "limited resources" cost to using time or choices, which would lead to more differentiation in the action patterns.

7. CONCLUSION

Highly interactive, high-resolution log file data from virtual performance assessments show promise for documenting in new ways what students know and can do. Data mining, machine learning and symbolic regression techniques are effective tools for analyzing and making sense from the time-based records and for relating those to both automated and human scoring artifacts. New psychometric challenges are emerging due to the dynamics, layered resolution levels, and complex patterning of actions with objects in virtual performance assessment spaces. Learning analytics analyses are helping uncover and articulate the relationship of time-event appraisals, visualization structures and resource utilization constraints on the psychometrics of virtual performance assessments.

REFERENCES

- Campanharo, A. S. L. O., Sirer, M. I., Malmgren, R. D., Ramos, F. M., & Amaral, L. a N. (2011). Duality between time series and networks. *PloS one*, 6(8), e23378. doi:10.1371/journal.pone.0023378
- Clarke-Midura, J., Dede, C., Norton, J. (2011). Next Generation Assessments for Measuring Complex Learning in Science. The Road Ahead for State Assessments. MA: Rennie Center for Education Research & Policy.
- Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Stastical Society. Series B (Methodological)*, 39(1), 1–38.
- Gigerenzer, G., Todd, P., & ABC, R. G. (1999). *Simple heuristics that make us smart* (p. 416). Oxford: Oxford University Press.
- Han, J., Cheng, H., Xin, D., & Yan, X. (2007). Frequent pattern mining: current status and future directions. *Data mining and knowledge discovery*, 15(1), 55–86. doi:10.1007/s10618-006-0059-1
- Ifenthaler, D., Eseryel, D., & Ge, X. (2012). *Assessment in game-based learning*. (D. Ifenthaler, D. Eseryel, & X. Ge, Eds.) (p. 461). Springer.
- 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.
- Miles, M., & Huberman, A. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks, CA: Sage Publications.
- Mislevy, R.J. Behrens, J.T., DiCerbo, K.E., Frezzo, D.C., & West, P. (2012). Three things game designers need to know about assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (eds.), *Assessment in game-based learning: Foundations, innovations, and perspectives*. New York: Springer.
- Schmidt, M., & Lipson, H. (2009). Symbolic regression of implicit equations. *Genetic programming theory and practice*, 7(Chap 5), 73–85.
- Tobias, S., & Fletcher, J. D. (2011). *Computer games and instruction*. Information Age Publishing.
- Witten, F., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (p. 524).
- Wolfram, S. (2002). *A new kind of science*. Champaign, IL: Wolfram Media.