Extending Semantically Enabled Virtual Environments for Training Assessment

Christian Greuel, John Murray, Maneesh Yadav

Affiliation:

SRI International, 333 Ravenswood Avenue, Menlo Park, California 94025, USA

Correspondence to: firstname.lastname@sri.com

Abstract: Educators and trainers at all levels are interested in deploying game-based environments and virtual environments as innovative educational tools and intelligent training systems, especially in application areas that involve challenging task activities. However, many development technologies are difficult to use, especially for those who are not specialists in building computer-based systems. In particular, the incorporation of lesson preparation and learner assessment into a computer-based learning program deserves close attention. This paper reports the lessons learned from several recent research and development projects, and offers some directions for new studies that build on this work. The core work of interest is Semantically Enabled Automated Assessment in Virtual Environments. This prototype task-training framework assesses learner performance within an instrumented virtual environment and provides contextual feedback to help improve skill acquisition. An authoring capability allows subject matter experts to create exercise solution models by demonstrating them directly in the environment. A review of additional research offers avenues for improving the existing exercise development and learner assessment capabilities.

One Sentence Summary: Semantically grounded virtual environments bridge the gap between the graphical representation of the training scene and the instructional models used to drive learner assessment, yet current research can be applied to improve the quality of this enabling capability.

1. Introduction

Virtual environments (VEs) are an increasingly attractive method for learning new procedural skills or refreshing one's existing skills, particularly in domains where similar training in the real-world would incur significant time, expense, and/or risk. SRI International (SRI) recently developed a prototype framework, Semantically Enabled Automated Assessment in Virtual Environments (SAVE) [1], for observing and assessing learner performance of an exercise within a training VE. We also investigated other research topics for key insights that indicate approaches for potential iterative improvements in the assessment veracity in the SAVE prototype.

This paper begins with a high-level overview of the SAVE framework and its approach to automated performance assessment by leveraging semantic overlays. We then describe applicable techniques from evidence-centered design (ECD) [2], a principled framework for the design, production, and delivery of educational assessments. From there, we discuss how an ECD-based toolkit can be deployed to integrate the roles and tasks of the various stakeholders involved, including designers, trainees, mentors, and analysts. We conclude by offering a potential framework for multimodal data analysis that synthesizes and produces results-based evidence to inform the next iteration of the lesson design.

2. Framework for Training Assessment in Virtual Environments

Maximizing the value from an educational VE requires a mechanism, such as direct observation by an instructor, to assess how well the learners have performed. Obviously, this approach is prohibitively expensive when large numbers of students are involved, and it becomes logistically impossible when the student population is geographically distributed, as is often the case with self-directed, computer-based learning systems. An automated mechanism that assesses learner performance can greatly reduce the cost of evaluating and improving training within VEs.

The SAVE framework, summarized in Figure 1, is designed to provide a free-play task trainer that can deliver meaningful feedback directly to the learner without the expensive physical presence of a live instructor. We developed this capability by integrating a unique method of automated assessment that leverages novel semantic overlay and graph-matching techniques to evaluate learner activity by comparing it to a previously defined exercise solution. The gold-standard reference solution is generated once by an instructional content author, simply by demonstrating the task in the VE.

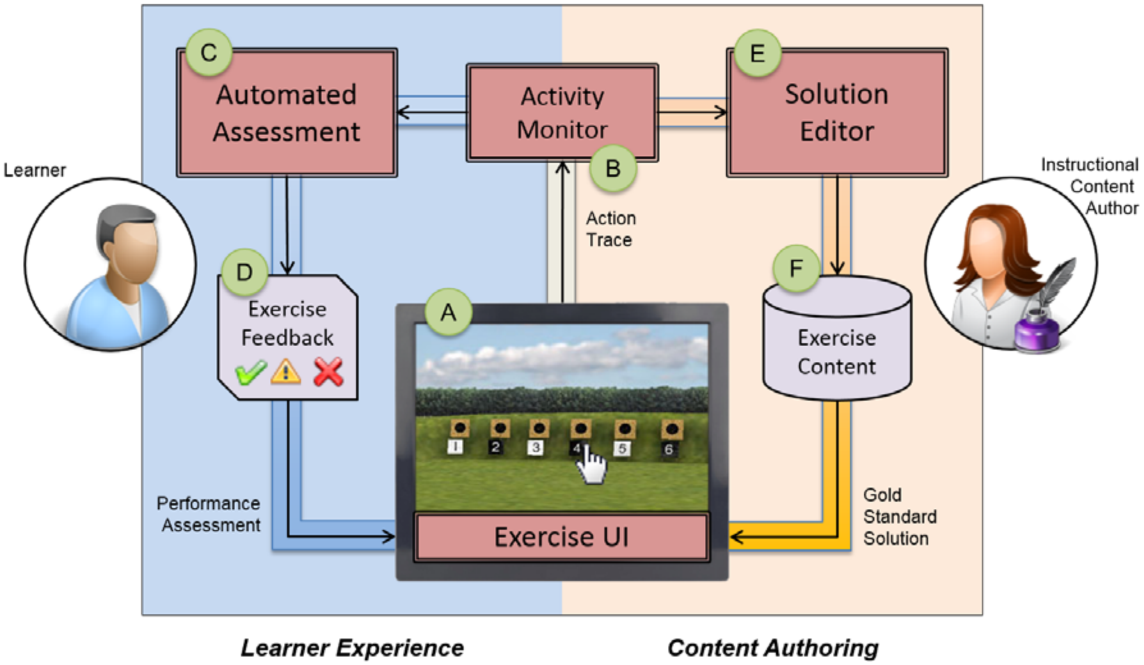

Figure 1. The SAVE framework tracks learner activity, assesses their performance, and provides contextual feedback to help improve skills. Assessment is facilitated by an authoring-by-demonstration technique that enables instructors to leverage the same user interface as the learner to specify core exercise solutions.

Providing these capabilities required the creation of a method for performing automated reasoning about user activity in the VE. The approach is to semantically annotate each object and its components in the VE and the actions that a user could perform upon them within an exercise user interface (EUI). When a learner pushes a switch or pulls a lever in such an annotated scene (A), a semantically defined action trace is generated (B) that clearly describes the activity. The collected sequence of such action traces provides an accurate narrative of learner performance upon which then reasoning is based.

After the learner completes an exercise, SAVE assesses the complete action trace to measure the learner's performance (C). Our automated assessment uses a flexible graph-matching technique first developed for training on the use of complex military planning software [3]. For each given exercise, a corresponding solution model is represented as a graph of all acceptable solutions. The assessment aligns the student response graph against the solution graph to locate an approximate match. Any misalignments against this match are considered student errors, which are reported back to the student as contextual feedback (D). This graph-based representation of multiple acceptable solutions enables effective performance evaluation while allowing the learner an amount of exploration.

An instructional content author generates the solution model for a given exercise in advance of the student training. In the same manner used to capture learner performance, the author demonstrates the proper solution directly in the EUI. This activity is collected as an action trace, just as it would be for a learner. However, instead of being used for assessment, the author's trace is used as the basis for an initial parameterized solution model. The author then generalizes the model by using a solution editor (E), which allows the author to relax certain constraints of the solution relating to either the steps (grouping, ordering, or optionality) or the variability of parameters (classification, enumeration, or range). The solution is stored with the associated exercise assets in a content repository (F). This authoring-by-demonstration technique provides an intuitive method for defining the core solution, while the generalization allows for expanding this to a broader set of acceptable solutions.

With this approach, SAVE assesses a learner's performance and provides contextual feedback to help improve skills and enhance understanding. In contrast to intelligent tutoring tools that assess "algorithmic" skills with single acceptable responses, SAVE addresses open-ended procedural skills, which have a range of acceptable solutions. SAVE's automated assessment is facilitated by content authoring tools that enable instructors to specify training exercises and solutions to those exercises.

3. Virtual Environment Analytics and Educational Assessment

Educational assessment is the process of measuring the knowledge, skills, and abilities achieved by learners in an instructional course or program. In recent years, the technological tools and methods used to create, field, and interpret such assessments have evolved significantly, so it is now feasible to integrate them with advanced VEs for sophisticated real-time assessment [4]. Such environments can be designed to support the administration of complex and realistic assessment tasks and the accumulation of direct evidence of learners' thinking, reasoning, or understanding [5]. Measurement and statistical models make possible the integration and interpretation of multiple pieces of information to support valid inferences about what learners know and can do.

However, few VE designers and developers have had much exposure to the methods and techniques, such as ECD, used by education professionals to create and validate learning assessment tools and instruments. Similarly, many assessment designers have a limited understanding of the data collection capabilities and analytical resources available to them within the VE development arena.

The emergence of social media technologies and multiplayer environments has promoted the shared acquisition of knowledge and the development of distributed skills, especially in the contexts of collaborative exploration and group learning. Because many educational assessment practices primarily focus on the individual student, as is currently the case with SAVE, they fail to account for knowledge-building and learning in a broader context. As assessment research considers the cultural shifts that arise from the emergence of a more participatory culture, new methods of designing and applying assessments to learning communities are needed.

To address this shortfall, inspiration may be drawn from the latest experimental procedures and statistical techniques that are used to characterize and evaluate real-world attributes and achievements from observations of gameplay in modern VEs. Of particular value are insights gained from the use of digital games in formal and informal educational settings, especially in their use as assessment and evaluation tools. It is also helpful to understand the process of dovetailing current classroom pedagogy techniques with new, game-based approaches to learner assessment and skills evaluation. Two recent meta-analyses have examined the effects of a comprehensive suite of educational simulations and games on learning outcomes [6, 7]. The focus of these analyses was to characterize the core design features and affordances in the game and simulations, and to compare their use in educational settings with traditional instructional methods.

4. Evidence-Centered Design

In addition to improving the results of educational VE’s, ECD practices are needed to help facilitate superior quality studies that can accurately assess the effects of computer-based learning interventions. As a rule, truly effective interventions are rare, and most studies are motivated by efforts to promote specific interventions, thus producing unreliable or biased results. However, common elements in the design and implementation of such studies could be encapsulated within the ECD framework. These elements include assessment authoring and delivery systems, evidence and task models, and student performance scoring tools. The resulting resources would facilitate the evaluation of educational products ranging from standardized testing to procedural task learning within VEs.

ECD principles and methods can be encapsulated into application platforms and toolkits that can be customized for specific educational and training contexts. For example, the Principled Assessment Designs for Inquiry (PADI) platform [8] is an implementation and extension of ECD that includes tools for incorporating national educational standards into experiment designs. While ECD techniques have been applied in traditional educational product development, our experience in VE design and assessment suggests that ECD tools need additional enhancement to support the VE design process, especially in the prototyping stage when feedback is most valuable.

Towards this end we recommend the development of software engineering tools that improve the design of educational VEs by enhancing communication across the diverse set of specialists needed in VE development (e.g., software engineers, course developers, graphic artists, and educational assessment experts). This process promotes effective workflows that allow assessment experts to specify and prioritize particular VE features that will maintain high-quality assessment validity. For example, close attention to the VE design elements of a crucial learner interaction can help differentiate between random trial-and-error attempts and the careful exploration of alternative solutions.

Development harnesses that interface with integrated development environments (IDEs), e.g., Eclipse, can be used to automate verification of product components, e.g. telemetry logging, which are often the keys to assessment validity. Assessment experts can also annotate code and game assets to ensure that correct experiments are run, while minimizing the communication burden on the software engineering team, where ad hoc measures and compromises are often required to meet product delivery deadlines.

5. Educational Virtual Environments: A Design-For-Assessment Strategy

The use of an ECD approach can integrate the activities of the various stakeholders involved in the research and development of digital educational products such as VE games and simulations. As illustrated in Figure 2, the stakeholders include:

- The VE design team, which includes the technical experts and domain specialists
- The study participants, which includes the student/learner population along with instructors and mentors, where appropriate
- The performance scoring team, who are responsible for evaluating the educational performance and achievements of the students/learners
- The data analysis team, who identify the participant behaviors and activities within the VE that are pertinent to the assessment process

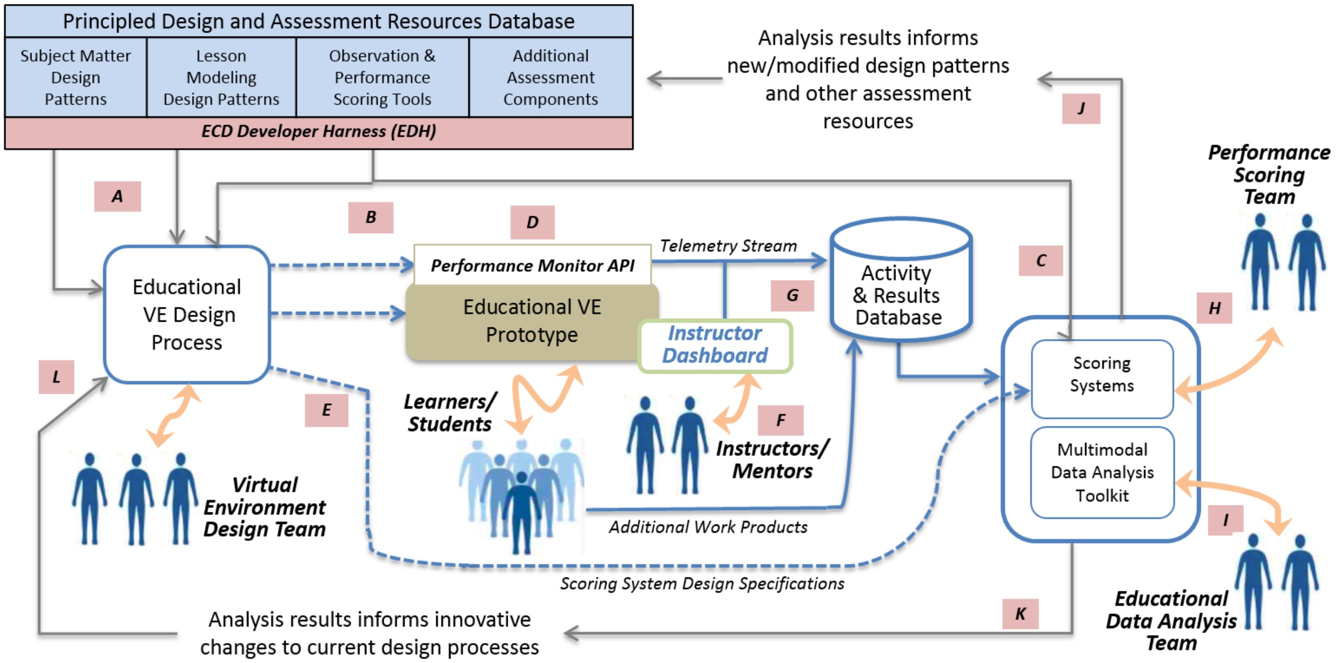
Principled Design and Assessment Resources Database
Subject Matter Design Patterns
Lesson Modeling Design Patterns
Observation & Performance Scoring Tools
Additional Assessment Components
ECD Developer Harness (EDH)
Analysis results informs new/modified design patterns and other assessment resources
J
A
B
D
C
Performance Scoring Team
Educational VE Design Process
Performance Monitor API
Telemetry Stream
Educational VE Prototype
G
Activity & Results Database
Instructor Dashboard
H
Scoring Systems
L
E
Learners/ Students
F
Instructors/ Mentors
Multimodal Data Analysis Toolkit
Virtual Environment Design Team
Additional Work Products
I
Scoring System Design Specifications
K
Educational Data Analysis Team
Analysis results informs innovative changes to current design processes

Figure 2. The design team uses the ECD Developer Harness as part of a design-for-assessment strategy in the development of an educational VE. Student performance in the environment is analyzed by the scoring team, whose findings are then used by the design team for the next iteration of their product.

An ECD Developer Harness (EDH) supports the development and use of digital VE games and/or simulations by the stakeholder population. The EDH includes a standard suite of software engineering systems and resources, including an IDE; application program interfaces (APIs); quality assurance (QA) scripts; and similar tools. In addition to ensuring alignment between the digital educational products and learning goals, the EDH enables designers and developers to map ECD-oriented design patterns to source code, assets (game art), and test results (results from unit, cognitive labs, usability, and summative tests) so that these mappings are persistent and can be disseminated. As a bridge between principled design and assessment resources and the VE developer outputs, the EDH ensures that appropriate design decisions map specific elements of the VE game or simulation to specific elements of the ECD knowledge ontology. Prior experience with paper-based systems similar to ECD suggests that the EDH could quadruple the productivity of multi-disciplinary design teams by halving development time and doubling the acceptance rates of educational VE products [4, 5].

In Figure 2, members of an educational VE design team use the EDH to draw from subject matter and lesson-modeling design patterns in the resources database (A). They also apply observation and scoring models (B) from the database as part of the product specifications. These models are also available for use in the results scoring system (C).

As the VE development work progresses from initial design to full implementation, the development team instruments the prototype with a suite of performance monitoring functions (D). These functions are software hooks that form the API component of the EDH. As part of the design process, the team also produces the specifications (E) that are required for the scoring system. These specifications are documented using the EDH and are provided directly to the scoring system.

The prototype game is deployed into the educational setting, where is it used by students under the guidance of their instructor or mentor (F). As the students work with the prototype, the performance monitor logs their interactions with the system and generates the telemetry stream that is recorded in the activity and results databases (G). At the same time, the dashboard provides instructor access to students' individual and group performance results. Additional student work products (worksheets, written reports, etc.) are also transferred to the database.

The performance scoring team uses the scoring system to apply rubrics and other scoring models (Bayesian networks, artificial neural networks, hidden Markov models) to student responses (H). Educational analysts (I) use a suite of multimodal analysis tools to examine students' in-game activities and confirm research hypotheses about learning performance. The results of these analyses are used to inform new or modified principled design patterns and resources (J) and define innovations and revisions to the current design team processes (K). The VE design team (L) uses the outcomes and findings to update the next iteration of the learning system.

6. Integrating Qualitative and Quantitative Analytics

The deployment framework and the use of the multimodal analysis tools allow instructors, developers, and assessment experts to explore and investigate the learner and instructor Activity and Results Database, as shown above. Figure 3 summarizes the exploration process, which links VE activity and observation data for multiple learner and instructor participants with individual and group behavioral models and characteristic patterns. This integrative approach draws on Actor Network Theory (ANT) [9, 10] to structure the complex links among the study participants (instructors and learners) and their relationships with the configurations, artifacts, and agents in the VE. This approach provides a powerful framework for expressing research hypotheses that seamlessly dovetail quantitative and qualitative analytical techniques.

Participant data from the Activity and Results Database is classified into four separate groupings:

- (A1) Communication among individuals, as captured either by in-world voice recordings or chat logging, or by local input device (e.g., keystroke) logging on the participant's platform
- (A2) Intrinsic and naturalistic behaviors and actions, such as in-world movements, posturing, attitude adoption, etc.
- (A3) Purposeful in-world determinations, such as selected choices among options, acceptance of risks, decisions that affect the overall direction of the lesson or VE experience
- (A4) Video streaming data or lab participant observations, captured by laboratory cameras, eye-tracking systems, or other psychophysical sensing systems

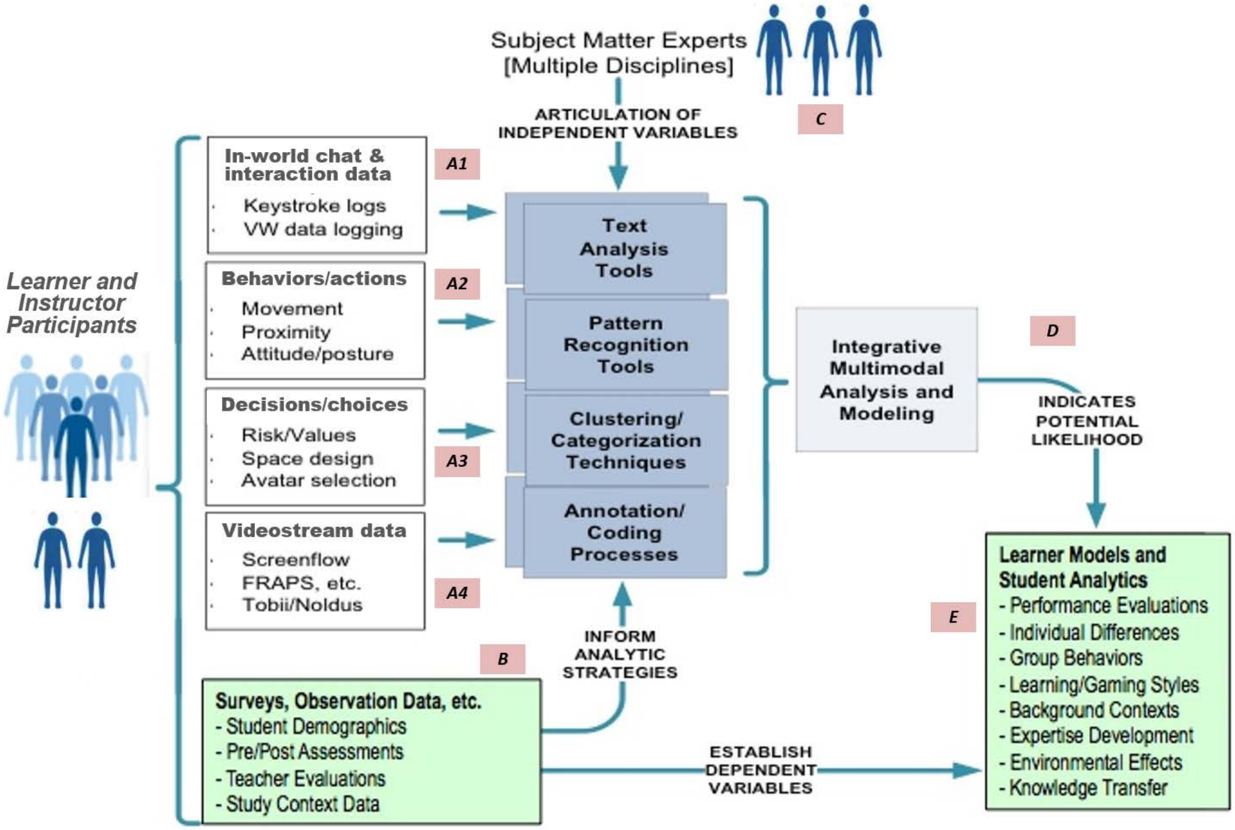

Figure 3. The multimodal analysis process integrates activity and observation data for multiple learners according to behavioral models and characteristic patterns. Gathered data is classified, processed, and analyzed to generate findings that associate real-world characteristics of participants with their virtual behaviors.

Information about participants' real-world demographics, educational evaluations, survey data, etc. is also gathered as part of the study (B). This information is used to identify strategies for understanding and interpreting the project data and provides dependent variables for the statistical analyses.

Developers can apply a broad suite of coding processes and analytical tools to the participant activity data streams, with decisions about the independent variables of interest guided by subject matter experts (C). In this context, team members pay particular attention to the specific affordances, modifiability, and limitations of the virtual world resources, inventory, and behaviors. In many cases, subject matter experts will adopt a more qualitative strategy to examining and explaining findings in this type of study. Such interpretive methods are often found among researchers in the arts and humanities, where ANT provides a productive means to link qualitative and quantitative research techniques.

With these techniques, the output of the multimodal analysis and modeling process generates findings that associate real-world characteristics of participants with their VE behaviors (D). For another recent project, these techniques were used to identify in-world behavioral indicators that differentiated between categories of game participants based on gender, age bracket, ethnic background, and socio-economic status, among others [11]. Additional findings provided insights into the propensity of participants to team up in pursuit of game objectives and their typical roles and behaviors within such groups (E).

We believe that there is considerable scope for applying the analysis insights – at the individual and the team levels – to the design and development of educational VEs for training activities involving sophisticated tasks, which can be provided to learners either singly or in groups.

7. Conclusions

The initial development work on SAVE has provided an underlying proof-of-concept prototype that demonstrates the practical use of VEs in educational and training applications. The technique of authoring-by-demonstration can allow non-technical subject matter experts to create valid exercise solutions to support the automated assessment of learner performance.

With this background, we have discussed several means to build on the basic SAVE implementation using the development and analytical frameworks that have been outlined. The design-for-assessment approach introduces an iterative experimental protocol for adapting ECD resources to an educational VE system. Such a process will help revise VE-specific resources and provide a broader reach for established techniques beyond traditional assessment-driven lesson and curriculum development.

A multimodal analysis framework provides data-based insights into learner strategies and behaviors, which can advise educational experience developers on potential innovations within the design process. In other arenas [11], we have found that this process offers a practical means of combining the values and hypotheses of qualitative researchers in the arts and humanities with the structured technical tools and techniques that are commonly used in the behavioral and natural sciences. Building on the success of the SAVE prototype, future performance-assessment capabilities will be able to integrate and collate the discreet actions of learners with their real-world characteristics and behavioral qualities.

References and Notes

1. C. Greuel, K. Myers, “Semantically-enabled automated assessment in virtual environments: project phase report, year 1 (final)” (PS-22266-TR-15-042, SRI International, Menlo Park, CA, 2015).
2. R. J. Mislevy, G. D. Haertel, Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*. **25(4)**, 6-20 (2006).
3. K. Myers, M. Gervasio, C. Jones, K. McIntyre, K. Keifer, Drill evaluation for training procedural skills. *Proceedings of 16th International Conference on Artificial Intelligence in Education*, 561-570 (2013).
4. G. D. Haertel *et al.* Design and Development of Technology-enhanced Assessment Tasks: Integrating Evidence-Centered Design and Universal Design for Learning Frameworks to Assess Hard-to-Measure Science Constructs and Increase Student Accessibility. Paper presented at the Invitational Research Symposium on Technology Enhanced Assessment, Washington DC, (2012).
5. T. P. Vendlinski, G. K. W. K. Chung, K. R. Binning, R. E. Buschang, Teaching rational number addition using video games: the effects of instructional variation (CRESST Report 808, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles, CA, 2011).
6. D. B. Clark, E. E. Tanner-Smith, S. Killingsworth, “Digital games for learning: a systematic review and meta-analysis,” SRI International, Menlo Park, CA (2014).
7. C. D’Angelo *et al*, “Simulations for STEM Learning: Systematic Review and Meta-Analysis,” SRI International, Menlo Park, CA (2014).
8. Information on Principled Assessment Designs for Inquiry (PADI) is on the website http://padi.sri.com .

9. B. Latour, “Where are the missing masses? The sociology of a few mundane artifacts,” in *Shaping Technology/Building Society: Studies in Sociotechnical Change*, W. E. Bijker, J. Law, Eds. (MIT Press, USA, 1992), pp. 225-258.
10. B. Latour, *Reassembling the Social: An Introduction to Actor-Network-Theory* (Oxford Univ. Press, New York, 2005).
11. J. Murray *et al*, “Reynard VERUS final report” (AFRL-RY-WP-TR-2012-0286, US Air Force Research Laboratory, Wright-Patterson AFB, OH, 2012).

Acknowledgments

This material is based upon work supported by the United States Government under Contracts No. W911QY-14-C-0023 and No. FA8650-10-C-7009. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Government.