Predicting Successful Inquiry Learning in a Virtual Performance Assessment for Science

Ryan S.J.d. Baker[1], Jody Clarke-Midura[2]

[1]Teachers College, Columbia University, New York, NY, USA
baker2@exchange.tc.columbia.edu

[2]Harvard Graduate School of Education, Cambridge, MA 02138
jody@post.harvard.edu

Abstract. In recent years, models of student inquiry skill have been developed for relatively tightly-scaffolded science simulations. However, there is an increased interest in researching how video games and virtual environments can be used for both learning and assessment of science inquiry skills and practices. Such environments allow students to explore scientific content in a more open-ended context that is designed around actions and choices. In such an environment, students move an avatar around a world, speak to in-game characters, obtain objects, and take those objects to laboratories to run specific tests. While these environments allow for more autonomy and choice, assessing skills in these environments is a more difficult challenge than in closed environments or simulations. In this paper, we present models that can infer two aspects of middle-school students' inquiry skill, from their interactive behaviors within an assessment in a virtual environment called a "virtual performance assessment" or VPA: 1) whether the student successfully demonstrates the skill of designing controlled experiments within the VPA, and 2) whether a middle-school student can successfully use their inquiry skill to determine the answer to a scientific question with a non-intuitive in-game answer.

Keywords: student modeling, skill modeling, inquiry learning, virtual performance assessment

1 Introduction

Over the last decades, the field of user modeling and adaptive personalization, and related research communities, have worked to extend student modeling methods from more well-defined domains and learning systems to more ill-defined situations. Highly successful approaches for well-defined tutoring systems within domains such as mathematics and physics [6,8], have been followed with steps to extend student modeling to learning systems in more ill-defined domains and/or for more ill-defined competencies [cf. 3, 7, 10, 13, 15, 16].

Specifically, student modeling has been extended to the study of student inquiry skill within simulations and microworlds. For example, Sao Pedro and colleagues have developed machine-learned models that can infer whether a student has the skills of designing controlled experiments and testing stated hypotheses, within scientific simulations that scaffold the process of selecting hypotheses, collecting data, and interpreting it to answer questions about the data [15,16]. In addition, Quellmalz and colleagues have developed knowledge-engineered models that infer student inquiry skill based upon well-scaffolded activities where students make observations, run experimental trials, and interpret data by answering questions about graphs and create concept maps [12].

Beyond these relatively structured scientific simulations, work has also extended modeling of students' scientific inquiry skills to less structured virtual environments, where avatars move around a large environment, acquiring objects of interest, interacting with non-player characters, and running experiments in a laboratory. Rowe and Lester used knowledge engineering to define the structure of a Dynamic Bayesian Network, which infers a learner's narrative knowledge, strategic knowledge, and content knowledge [13]. This DBN was shown to accurately predict student responses on content test questions, after use of the virtual environment. Sil and colleagues used machine learning to develop models which can replicate human judgments about the quality of an essay where the student justifies their hypotheses for why a phenomenon is occurring in a virtual environment [18]. Their models utilize information both from the linguistic features of the essay, and from features of the student's interaction with the virtual world.

In this paper, we present a model that can infer whether a student will successfully use their inquiry skill to 1) determine the answer to a scientific question with a non-intuitive in-game answer, and 2) successfully justify their claim based on causal evidence from within the game. We predict these aspects of inquiry skill using data from the student's interactions with an open-ended virtual environment, discussed in the following section, where students move around a virtual world, physically collecting data, talking to non-player characters, and running experiments in a laboratory. Only data from students' interactions with the virtual environment are included in the factors used to make predictions. As such, we both infer a key type of inquiry skill (whether the student can justify their claim) and whether the student has acquired content knowledge from their inquiry (whether the student produces the correct answer to the scientific question), doing so in the context of an open-ended virtual environment.

2 Virtual Performance Assessments

We conduct this research in the context of the Virtual Performance Assessment project at the Harvard Graduate School of Education. This project is developing and studying the feasibility of immersive virtual performance assessments to assess scientific inquiry of middle school students as a standardized component of an accountability program (see http://vpa.gse.harvard.edu). The goal is to provide states with reliable and valid technology-based performance assessments linked to state and national academic standards around science content and inquiry processes.

The virtual performance assessments are designed in the Unity game development engine [19]. The immersive nature of the three-dimensional (3D) environment allows for the creation and measurement of authentic, situated performances that are characteristic of how students conduct inquiry (see [11]). Students have the ability to walk around the environment, make observations, gather data, and solve a scientific problem in context. Further, these environments enable the automated, invisible, and non-intrusive collection of students' actions and behaviors during the assessment play.

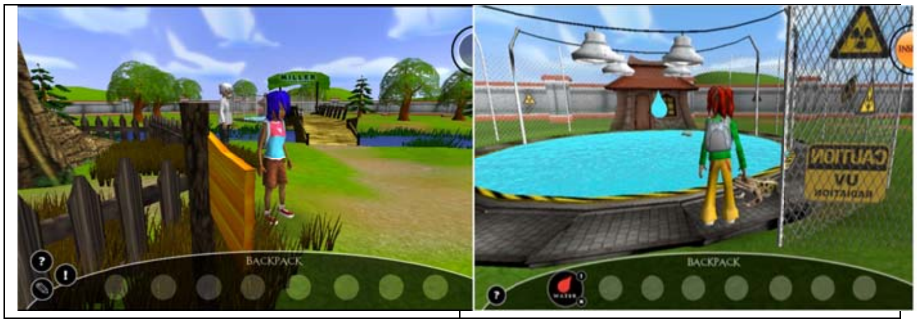

Fig. 1. Screen shots of the Virtual Performance Assessments (VPA)

As seen in Figure 1, the assessment has the look and feel of a videogame, yet places students at the center of a scientific problem that they have to solve. Thus, the goal is to develop assessments that measure students' science learning *in situ*.

In the Virtual Performance Assessment studied within this paper, students must learn why the frogs at one farm have six legs, and can choose between a set of hypotheses, including parasites (the correct hypothesis), pollution, pesticides, genetic mutation, and space aliens. They can study frogs and water from different farms in order to collect evidence related to their hypotheses.

Students log into each assessment via a web browser. Once logged in, they work individually through the scientific problem. The assessments measure various aspects of students' inquiry skills, including:

- Student develops a causal explanation of what is happening in the virtual world that culminates in: 1) a claim about the phenomena, 2) the evidence (either empirical or observations), and 3) reasoning that links claims with evidence.
- Student gathers data that help explain or provide evidence to justify the claim being made.

As mentioned above, in the course of interacting with the VPA, students perform many actions that are recorded and stored as log data. This creates a large data set showing what students have viewed, collected and used in their experiments. Unlike traditional assessments, students are free to choose which tasks to do and in what order. This simulates a real world environment where there are choices, wrong paths and extraneous information.

3 Data Set

The data analyzed in this study were produced by 1,985 middle school students. These students were in grades 7-8, and were 12-14 years old. These students used the Virtual Performance Assessment within their science classes, spread across 40 teachers and 138 classrooms across the Northeastern and Midwestern US, and Western Canada.

The 1,985 students each used the Virtual Performance Assessment until they had completed analysis and produced a final answer, spending an average of 29 minutes and 29 seconds (SD = 14 minutes and 30 seconds) in the environment. As students used the VPA, their actions within the software were logged, including the type of action (for instance, moving between regions of the VPA, picking up and inspecting objects, running laboratory tests on objects, reading informational pages, and talking to non-player characters), the location of the action (in terms of a set of game regions), the object being manipulated (including specific game objects such as yellow frogs, specific informational pages, and non-player characters being interacted with), details of the interaction (such as which tests were run in the laboratory, or which topics were discussed with a non-player character), and the time stamp.

These log files were further distilled for analysis, producing a set of 48 semantically meaningful features that could be analyzed further. These features were of the following types:

- The number of times a student went to specific types of locations (e.g. farms to collect evidence, the laboratory to run tests) and the ratio between these values (e.g. how many farms did the student visit per trip to the laboratory)
- What percentage of time was spent in specific locations?
- How full the student’s backpack became, both including repeats (e.g. picking up two green frogs counts as two objects), and not including repeats (e.g. two green frogs counts as one object) – both maximum fullness and average fullness over time were calculated
- How many times the student brought specific objects to the lab, including a count of how many distinct non-sick frogs were brought to the lab

- Number of tests run in the lab, both simultaneously, and separately, including and not including tests run multiple times
- Maximum degree of object coverage of a lab test across all tests so far (where full coverage would involve running a test on every object that test can be run on)
- How many tests were run on specific objects of interest (e.g. how many tests were run on the six-legged frog? How many tests were run on the frogs that were not ill?), both overall, and for specific tests
- How long did students pause (possibly to self-explain [cf. 17]) after running tests? (Average, Sum, and Standard Deviation)
- How many informational pages did student read, both overall and for specific pages of interest (e.g. the pages on the key hypotheses being investigated)? How many times did students read those pages?
- How long did students spend reading informational pages, both overall and for specific pages of interest? (Standard deviation was also computed for this metric, across all pages)

3.1 Dependent Measures

Two measures were predicted from students' behavior within the VPA: 1) the correctness of their final conclusion as to why the frogs had six legs and 2) their skill in designing causal explanations (DCE) for why that claim was correct.

The measure of the correctness of the student's final conclusion was based on whether the student's final conclusion was fully correct – e.g. did the student select at the end of the activity that the 6-legged frogs were caused by parasites, or did the student select one of the other potential hypotheses.

Designing causal explanation is defined as the student's ability to support their claim or conclusion with evidence. Most of the evidence in the VPA was consistent with parasites being the cause of the 6-legged frogs. Three of the four other claims had at least some evidence consistent with the claim: pollution, pesticides, and genetic mutation (there was no evidence in favor of space aliens), but there was other evidence against these claims. While even the non-causal data was strong enough to show that these claims were unlikely to be the cause, students were given partial credit if they provided supporting evidence for these claims. The students who are most successful at inquiry will be able to make a causal explanation and support it with causal inference; less successful students may fail at doing this, still demonstrating the ability to support a claim but not demonstrating an ability to fully distinguish causal and non-causal 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. This evidence on DCE was aggregated across indicators into a single measure. The mean DCE was 50%, with a standard deviation of 23.33%.

4 Detector of Designing Causal Explanations

The detector of students' ability to design causal explanations was set up as a linear regression, as the metric for DCE was numerical. Linear regression was implemented using the M5' variable selection procedure [21] in RapidMiner 4.6 [9], using Leave One Out Cross Validation (LOOCV), at the student level (which was the overall level of analysis). Linear regression was chosen as a relatively conservative algorithm, with a relatively low probability of over-fitting. Correlation was used as the goodness metric.

The final model achieved a cross-validation correlation of 0.531 to the student's success in designing causal explanations, comparable to the success of Bayesian Knowledge Tracing models attempting to predict post-test performance in more tightly-constrained intelligent tutoring systems domains such as genetics problem-solving [cf. 1].

When trained upon all data, the final model was as follows:

```
DCE =
- 0.165 * Maximum number of items in backpack (including repeats)
+ 0.322 * Maximum number of items in backpack (not including repeats)
- 0.656 * Average number of items in backpack (including repeats)
+ 0.651 * Average number of items in backpack (not including repeats)
+ 3.483 * Maximum degree of coverage for a lab test
- 5.120 * Percentage of time student spent at farms
- 0.644 * Ratio between trips to lab and trips to farms (lab trips
          divided by farm trips)
- 0.197 * Number of different (types of) non-sick frogs student took to
          the lab at the same time
+ 0.542 * Did the student ever run a blood test on the six-legged frog
          {0,1}?
+ 0.714 * Did the student ever run a blood test on a non-sick frog
          {0,1}?
- 0.657 * Did the student ever run a genetic test on a non-sick frog
          {0,1}?
- 0.834 * Did the student ever run a water test on farm water {0,1}?
- 1.137 * Did the student ever run a water test on lab water {0,1}?
+ 1.372 * Number of times student took lab water to the lab
+ 0.044 * How long, on average, did students spend reading information
          pages? (average per read)
```

```
- 0.025 * Standard deviation of time spent reading information pages
          (per read)
+ 0.004 * How long, in total, did student spend reading information page
          on pollution
+ 0.009 * How long, in total, did student spend reading information page
          on parasites
- 0.563 * Total number of times student accessed information page on
          space aliens
+ 0.799 * Total number of times student accessed information page on
          parasites
+ 9.153
```

The general complexity of this model indicates that many factors are associated with a student's success in designing causal explanations. It is worth noting that, due to the complexity of this model, some features may have negative coefficients when incorporated into the overall model, while having a positive correlation when taken in isolation (suggesting that they are only negative when other features are taken into account). In fact, all of the features with a negative coefficient in the model actually have a positive correlation to DCE when taken individually, with the single exception of the percentage of time the student spent at farms. However, no features with a positive coefficient in the model actually have a negative correlation to DCE when taken individually.

Taken individually, the feature most strongly correlated with DCE is the total number of times the student accessed the information page on parasites, which had a non-cross-validated correlation of 0.488 to DCE. Note that non-cross-validated correlations typically have substantially higher values than cross-validated correlations (in other words, it is not correct to assume that this single feature accounts for almost all of the total variance of the cross-validated model). This relationship indicates that student ability to design causal explanations is strongly connected with other information-seeking behaviors.

Other features particularly correlated with DCE when considered individually include: the maximum degree of coverage for a lab test (r=0.301), the percentage of time spent at farms (r= -0.265), whether the student conducted blood tests on 6-legged frogs (r=0.231) or other frogs (r=0.227), whether the student conducted water tests on lab water (r=0.233) or farm water (r=0.201), whether the student brought lab water into the lab (r=0.241), whether the student conducted a genetic test on non-sick frogs (r=0.209), and the maximum number of distinct non-sick frogs brought into the lab (r=0.216). The reasonably high correlation seen for a range of features, and the complexity of the eventual model, indicate that there are a number of indicators which have some degree of independent prediction of DCE.

5 Detector of Correct Final Conclusion

The detector of whether the student's final conclusion was correct was set up as a binary classification problem. A small number of algorithms were attempted that had been successful on similar problems: J48 Decision Trees, JRip Decision Rules, Step Regression, and K*. The most successful algorithm was JRip Decision Rules [2], a rule induction algorithm based on information gain. JRip was implemented in RapidMiner 4.6 [9], using Leave One Out Cross Validation (LOOCV), at the student level (which was the overall level of analysis). Kappa and A' were used as the goodness metrics. Kappa assesses the degree to which a detector is better than chance at identifying which clips involve the correct conclusion. A Kappa of 0 indicates that the detector performs at chance, and a Kappa of 1 indicates that the detector performs perfectly. For example, a Kappa of 0.31 would indicate that the detector is 31% better than chance. A' is the probability that the algorithm will correctly distinguish an example of a correct conclusion from an example of an incorrect conclusion. A' closely approximates the area under the ROC curve in signal detection theory, also referred to as AUC ROC. A model with an A' of 0.5 performs at chance, and a model with an A' of 1.0 performs perfectly. The implementation of A' on our webpage (http://www.columbia.edu/ ~rsb2162/ edmtools.html) was used, as the implementation in RapidMiner 4.6 makes overly optimistic estimations in cases where several data points have the same confidence, a common case for JRip.

The final model achieved a cross-validated Kappa of 0.548 in predicting whether the student's final conclusion was correct, and an A' of 0.79 in doing so. This value for Kappa is comparable to detectors of science inquiry skill developed in much more constrained scientific simulation environments [e.g. 15, 16].

The final detector was as follows:

1. IF the student spent at least 66 seconds reading the parasite information page,
   THEN the student will obtain correct final conclusion (confidence = 81.5%)
2. IF the student spent at least 12 seconds reading the parasite information page
   AND the student read the parasite information page at least twice
   AND the student spent no more than 51 seconds reading the pesticides information page,
   THEN the student will obtain correct final conclusion (confidence = 75.0%)
3. IF the student spent at least 44 seconds reading the parasite information page
   AND the student spent under 56 seconds reading the pollution information page,
   THEN the student will obtain correct final conclusion (confidence = 68.8%)
4. OTHERWISE the student will not obtain correct final conclusion (confidence = 89.0%)

It is worth noting that this detector relies solely upon students' time spent reading specific information pages. The most important thing (according to this model) is whether a student spent substantial time reading pages about the correct final conclusion, that parasites can cause frogs to have six legs. It can also be seen that spending too much time reading pages linked to incorrect hypotheses can be indicative that the student will eventually choose the wrong final claim.

5.1 Other Features Associated with Correct Final Conclusions

It is noteworthy that a student's behavior beyond reading informational pages – acquiring objects, experimenting in the lab, and talking to non-player characters – is not incorporated into the model given above. This finding indicates that the informational pages are highly important for learning about the domain; however, the finding does not necessarily indicate that the other activities are useless, but indicates that the other activities lead to correct understanding only if combined with time spent reading specific information pages. We can study whether the other features of student behavior in the virtual environment also matter, by re-running the model without using features regarding the student's use of the information pages.

When we do so, again using JRip, the final model achieves a cross-validated Kappa of 0.118 in predicting whether the student's final conclusion was correct, and an A' of 0.56 in doing so. These values are much worse than the goodness achieved when time spent reading specific information pages is taken into account – but these values are still above chance.

The resultant detector is as follows:

1. IF the student achieves coverage for a lab test of 80% or higher at least once,
   AND the student brings 3 or more types of non-sick frogs into the lab at least once,
   AND the students brings an average of 1.94 or more different types of non-sick frogs into the lab together across lab trips,
   THEN the student will obtain correct final conclusion (confidence = 69.0%)
2. IF the student achieves coverage for a lab test of 100% at least once,
   AND the student conducts no more than 8 total sets of lab tests,
   THEN the student will obtain correct final conclusion (confidence = 70.6%)
3. OTHERWISE the student will not obtain correct final conclusion (confidence = 79.2%)

As such, we can see that when time spent reading pages is no longer taken into consideration, the remaining features that are predictive of a correct final conclusion involve how dense the student's testing procedures are – e.g. how many tests the student ran at the same time, and how many total testing episodes there were. Tests run on non-sick frogs are particularly meaningful for this, suggesting that it is important for the student to see what features are common and different among frogs that are not sick. Simultaneous tests are apparently relatively more useful for enabling students to integrate different results, perhaps because of the difficulty of remembering and mentally comparing tests that are more spread out in time. However, the benefits of these behaviors are relatively minor for obtaining correct conclusions, compared to student time spent processing the information pages.

6 Discussion and Conclusions

In this paper, we have presented models that can infer students' ability to design causal explanations, and whether they will successfully conduct science inquiry to the degree that they obtain a correct final claim about the phenomenon being studied. Models of scientific inquiry have been developed for other domains and learning systems in recent years, but often in fairly well-scaffolded environments [e.g. 12, 15, 16]. In this paper, we extend these methods to model these constructs within a more open-ended inquiry game environment, where students physically collect data, talk to non-player characters, and run experiments in a laboratory. Rowe and Lester have successfully predicted student content knowledge from use of such an environment [13]; we extend this work to also infer students' ability to design causal explanations, and to infer student learning of the key question in the simulation. As such, this work brings together the goals of inferring content learning and inquiry skill with the goal of embedding student inquiry learning in a lightly-scaffolded game-like environment.

The detectors presented in this paper are fairly coarse-grained in nature. They assess the overall success of student inquiry across about a half hour of student behavior. As such, they assess inquiry at a larger time-scale than many past approaches [e.g. 12, 13, 15, 16]. One advantage to the finer-grained assessment used in other approaches is that it more directly affords real-time intervention (e.g. if you know that a student has failed to demonstrate inquiry skill in the last 3 actions, you can intervene right away).

However, it is challenging to assess at this time-scale in more open-ended virtual environments, particularly for the types of aggregate skills and learning studied here. The lack of a key action having occurred so far (such as the student reading the informational page on parasites) does not imply that the key action will never occur. Unless a student's behavior is clearly unrelated to the learning goals [e.g. 14, 22], a student may be on an appropriate path but exploring some aspect of the simulation relatively more thoroughly than other students. As shown in the models above, relatively brief actions can have a disproportionate impact on overall success, and these actions are not required to have occurred before a specific point.

However, the coarse-grained nature of these models does not preclude them models from being a useful tool for improving the VPA. In its current design, the VPA can infer key aspects of a student's inquiry skill after a half hour: whether the student was successful at obtaining the correct final conclusion and in designing casual explanations. The models presented in this paper can be used to assess a student’s zone of proximal development as well: the difference between what the student can do without scaffolding, and what the student can do with help [e.g. 20]. For students who do not succeed at obtaining the correct final conclusion or in designing casual explanations, the model produced in this paper can be analyzed to select adaptive feedback. Each of the model features can be computed, and weighted as in the linear regression model above, in order to identify which feature(s) are most strongly associated with the student’s inquiry failure. Then the student can be given feedback related to these model features – for instance, a non-player character could say “Hmm, I’m not sure your causal explanation was as good as it could have been. Did you ever run a water test on lab water?” As such, it would be possible to assess not just whether the student’s initial inquiry is perfect, but how much feedback and help are needed to obtain fully correct answers. Approaches along these lines have increased the predictive power of assessment models in other domains [5]. The information from the model could also be used to help students reflect on their inquiry process after completing the VPA, to consider which of their decisions were effective and ineffective, towards promoting the student developing deeper meta-cognition about their inquiry skill.

The existing models could also be used to provide more in-the-moment feedback, by analyzing the internal features of the models. Within this environment, it is difficult to be certain that a desired action that has not yet occurred will never occur. But it is less difficult to identify specific actions as problematic based on the internal features of the models. For example, the more time a student spends reading information pages on space aliens, the less likely they are to successfully design causal explanations. A student who is spending considerable time on this page could be prompted by a non-player character to consider other pages, or to think about what evidence could support the claim that space aliens are causing farm frogs to have six legs. As such, it becomes possible to use these models – which are fairly coarse-grained in general – for adaptive personalization during learning.

In the long-term, developing student models for more loosely scaffolded learning environments will enable the methods and impacts of student modeling and adaptive personalization to apply to a wider range of learning situations. As this development occurs, the potential of individualized learning to improve student outcomes will expand to a greater range of interactive learning environments.

Acknowledgements. The research presented here was supported by the Bill and Melinda Gates Foundation. We also thank Chris Dede for his support and suggestions, and Michael Sao Pedro and one of the anonymous reviewers for their suggestions.

7 References

1. Baker, R.S.J.d., Gowda, S., Corbett, A.T.: Towards Predicting Future Transfer of Learning. In: Proceedings of the 15th International Conference on Artificial Intelligence in Education, 23-30. (2011)
2. Cohen, W.: Fast Effective Rule Induction. In: Proceedings of the Twelfth International Conference on Machine Learning. (1995)
3. Dragon, T., Woolf, B., Murray, T.: Intelligent Coaching for Collaboration in Ill-Defined Domains. In: Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED 2009), 740–742. (2009)
4. Hanley, J., McNeil, B.: The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. Radiology, 143, 29-36. (1982)
5. Feng, M., Heffernan, N.T., Koedinger, K.R.: Addressing the assessment challenge in an online system that tutors as it assesses. User Modeling and User-Adapted Interaction: The Journal of Personalization Research, 19 (3), 243-266. (2009)
6. Gertner, A., VanLehn, K.: Andes: A Coached Problem Solving Environment for Physics. In: Proceedings of the 5th International Conference on Intelligent Tutoring Systems, 133-142. (2000)
7. Graesser, A.C., Chipman, P., Haynes, B.C., Olney, A.: AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. IEEE Transactions on Education, 48, 612-618. (2005)
8. Koedinger, K. R., Corbett, A. T.: Cognitive tutors: Technology bringing learning sciences to the classroom. In: R. K. Sawyer (Ed.), The Cambridge handbook of the learning sciences. Cambridge University Press, New York, NY. (2006)
9. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler, T.: YALE: Rapid Prototyping for Complex Data Mining Tasks. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), 935-940. (2006)
10. Mitrovic, A.: An Intelligent SQL Tutor on the Web. International Journal of Artificial Intelligence in Education 13, 2-4, 173-197. (2003)
11. National Research Council: A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. Washington, DC: National Academies Press. (2011)
12. Quellmalz, E.S., Timms, M.J., Silberglitt, M.D., Buckley, B.C.: Science Assessments for All: Integrating Science Simulations Into Balanced State Science Assessment Systems. Journal of Research in Science Teaching, 49 (3), 363-393. (2012)
13. Rowe, J., Lester, J.: Modeling User Knowledge with Dynamic Bayesian Networks in Interactive Narrative Environments. In: Proceedings of the Sixth Annual Artificial Intelligence and Interactive Digital Entertainment

Conference (AIIDE-2010), 57-62, Palo Alto, California. (2010).

14. Rowe. J., McQuiggan, S., Robison, J., Lester, J.: Off-Task Behavior in Narrative-Centered Learning Environments. In: Proceedings of the 14th International Conference on Artificial Intelligence in Education, 99-106. (2009)
15. Sao Pedro, M., Baker, R., Gobert, J., Montalvo, O., Nakama, A.: Leveraging Machine-Learned Detectors of Systematic Inquiry Behavior to Estimate and Predict Transfer of Inquiry Skill. User Modeling and User-Adapted Interaction: The Journal of Personalization Research, 23 (1), 1-39. (2013)
16. Sao Pedro, M. A., Baker, R.S.J.d., Montalvo, O., Nakama, A., Gobert, J.D.: Using Text Replay Tagging to Produce Detectors of Systematic Experimentation Behavior Patterns. In: Proceedings of the 3rd International Conference on Educational Data Mining, 181-190. (2010)
17. Shih, B., Koedinger, K.R., Scheines, R.: A response time model for bottom-out hints as worked examples. In: Proceedings of the 1st International Conference on Educational Data Mining, 117-126. (2008)
18. Sil, A., Shelton, A., Ketelhut, D.J., Yates, A.: Automatic Grading of Scientific Inquiry. In: Proceedings of the NAACL-HLT 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-7). Montreal, Quebec, Canada. (2012)
19. Unity Technologies. Unity Game Engine. (2010)
20. Vygotsky, L.: Mind in Society. Cambridge, MA: Harvard University Press. (1978)
21. Wang, Y., Witten, I.H.: Induction of Model Trees for Predicting Continuous Classes. In: Proceedings of the European Conference on Machine Learning. (1997)
22. Wixon, M., Baker, R.S.J.d., Gobert, J., Ocumpaugh, J., Bachmann, M. WTF? Detecting Students who are Conducting Inquiry Without Thinking Fastidiously. In: Proceedings of the 20th International Conference on User Modeling, Adaptation, and Personalization (UMAP 2012), 286-298. (2012)