What Different Kinds of Stratification Can Reveal about the Generalizability of Data-Mined Skill Assessment Models

Michael A. Sao Pedro
Worcester Polytechnic Institute
100 Institute Rd.
Worcester, MA 01609 USA
(508) 831-5000
mikesp@wpi.edu

Ryan S.J.d. Baker
Teacher's College, Columbia
525 W. 120th St., Box 118
New York, NY 10027 USA
(212) 678-8329
ryan@educationaldatamining.org

Janice D. Gobert
Worcester Polytechnic Institute
100 Institute Rd.
Worcester, MA 01609 USA
(508) 831-5000
jgobert@wpi.edu

ABSTRACT

When validating assessment models built with data mining, generalization is typically tested at the *student-level*, where models are tested on new students. This approach, though, may fail to find cases where model performance suffers if other aspects of those cases relevant to prediction are not well represented. We explore this here by testing if scientific inquiry skill models built and validated for one science topic can predict skill demonstration for new students and a new science topic. Test cases were chosen using two methods: student-level stratification, and stratification based on the amount of trials ran during students' experimentation. We found that predictive performance of the models was different on each test set, revealing limitations that would have been missed from student-level validation alone.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: *Data Mining*; J.1 [**Administrative Data Processing**]: *Education*; K.3.1 [**Computer Uses in Education**]: *Computer-assisted instruction (CAI), Computer-managed instruction (CMI)*

General Terms

Measurement, Reliability

Keywords

Science Microworlds, Science Simulations, Science Inquiry, Automated Inquiry Assessment, Educational Data Mining, Learning Analytics, Validation, Generalizability, User Modeling

1. INTRODUCTION

Data mining/learning analytics is a powerful approach for building predictive models ("detectors") that determine if a student is engaging in a particular behavior (e.g. [1], [2]), and models that assess whether students demonstrate somewhat ill-defined skills within interactive learning environments (e.g. [3], [4]). Building models using data mining makes validation of those models easier, because processes like cross-validation exist to estimate how well they will generalize to new students and tasks not used to build them. Such estimates are important because they can provide assurance that the behavior / assessment models will correctly identify students who lack skill or engage in undesirable learning behaviors, enabling the system to provide accurate, real-time feedback [5]. Within open-ended interactive environments, the estimates can also assure that models that detect skill demonstration can be reused for new tasks/domains and students, paving the way for reliable, scalable performance-based assessment.

In educational domains, validation is often done at the *student-level,* where models are built using one group of students' data and tested on new students whose data were not in model construction [6], [1]. This ensures models will remain accurate when applied to completely new students. It is possible, though, that this method may fail to identify specific instances when models do not predict adequately, particularly if some other aspect of those cases, other than the student, is not taken into account. We explore this topic here in the context of determining how well two data-mined models that assess whether students demonstrate scientific inquiry skills, built and validated for one science topic [4], [7], can predict demonstration of those skills for a new science topic and a new set of students. Few papers have tested model effectiveness on different topics ([1] is an exception), but validating at this level is essential if models will be used beyond the topics in which they were originally developed.

In our approach, we first take this new topic and student sample, and construct a test set stratified at the student level, where students are equally represented in the test set. When doing this, we find that there is an imbalance in the nature of behaviors demonstrated by students. In particular, there is an imbalance in the number of trials collected by students in this set, a factor which could influence predictive performance of our models (cf. [7]). To address this, we construct a second test set, this time stratifying over the number of trials, to ensure a greater balance in student behaviors. We show that utilizing this different kind of stratification can unveil a different performance profile than conducting student-level validation alone, revealing new insights on the predictive limitations of the models.

2. PRIOR WORK BUILDING INQUIRY SKILL MODELS

In [4], [7], we developed data-mined models that assess students' demonstration of scientific inquiry process skills. Skills are assessed as students conduct inquiry within Inq-ITS activities (formerly known as Science Assistments [8]). This environment aims to automatically assess, scaffold, track, and provide real-time feedback on students' scientific inquiry skills. Inquiry is assessed as students explore within interactive simulations and use inquiry support widgets that facilitate experimentation.

Inq-ITS activities are performance assessments of inquiry skills. The actions students take within the simulation and work products they create are the bases for assessment [8]. This paper focuses on assessment of two process skills, designing controlled

LAK '13, April 08 - 12 2013, Leuven, Belgium

experiments and testing stated hypotheses. Briefly, students design controlled experiments when they generate data that make it possible to determine what the effects of independent variables (factors) are on outcomes. They test stated hypotheses when they generate data that can support or refute an explicitly stated hypothesis. Since these are process skills, students are assessed based on the actions they take while collecting data.

To build data mined assessment models, we employed text replay tagging of log files [9] to generate labels from which, in part, models were derived. A text replay is a "chunk" of student actions in text format that contains information enabling a human coder to identify what occurred in that sequence of actions. For our domain, text replays leverage human judgment to identify whether students demonstrate inquiry skill. These labels are then used as "ground truth" for whether or not students demonstrate a skill, and subsequently for building and validating detectors.

Prior to this paper, we have validated models for these skills for one physical science topic, Phase Change, and one student sample [4], [7]. One goal of this paper is to determine if these models can be applied to a new Physical Science topic, Free Fall, and to a new sample of students. As such, we first present a high-level description of the text replay tagging process within the context of assessing the two skills for Phase Change. A fuller description of the distillation process appears in [4], and of the model construction approach in [7].

2.1 Phase Change Activities

The Phase Change activities [8] aim to promote understanding about the melting and boiling properties of ice. Students learn about the topic by engaging in semi-structured scientific inquiry activities with a simulation. First, students are given an explicit goal to determine if one of four factors affects various measurable outcomes (e.g. melting or boiling point). Students then formulate hypotheses, collect and interpret data, and warrant their claims to address the goal. As mentioned, we developed data-mined models that infer whether students design controlled experiments and test stated hypotheses [4], [7]. These skills are demonstrated when students collect data to test their hypotheses in the "experiment" task. As such, we describe this task in more detail.

In the "experiment" task, students are shown the Phase Change simulation, graphs that track changes of the substance's temperature over time, and a table that captures all the data they collected thus far. They experiment by collecting data that aim to support or refute their hypotheses. For space reasons, we do not include a visual of this interface (it can be seen in [8]), but it is similar to the Free Fall interface shown in Figure 1. Students collect data (trials) by changing the simulation's variable values, and running, pausing and resetting the simulation. Next, we describe how these interactions were distilled and tagged with skill demonstration.

2.2 Distilling Raw Logs and Building Clips

Interaction data was collected from 148 middle school students who conducted inquiry within four Phase Change activities. All students' fine-grained actions were timestamped and recorded by the system. These actions included: interactions with the inquiry support widgets, interactions with the simulation including changing simulation variable values and running/pausing/resetting the simulation, and transitioning between inquiry tasks.

Contiguous sequences of these actions were then segmented into *clips*. A clip contains all the actions necessary for a human coder to identify (label) whether or not students demonstrate the inquiry skills. For our domain, a clip has all actions taken while formulating hypotheses ("hypothesize" actions) and collecting data ("experiment" actions) in an activity. Clips are the grain-size at which data collection skill is labeled, and in turn, student performance is assessed. From there, models describing what it means to demonstrate skill were trained and validated based on (1) labels indicating whether students demonstrated skills within clips, and (2) a set of features that summarize clips.

Humans determine whether or not students demonstrate skills within clips, by labeling *text replays* of clips with one or more tags (see [4] for an example of a text replay). For us, text replays highlight hypothesis creation and experimentation processes. This enables human to more easily identify skill demonstration. Clips are tagged as designing controlled experiments, testing stated hypotheses, both, or neither. In this prior work, we achieved acceptable levels of inter-rater reliability in labeling clips [4].

With clip labels in place, features were defined next that summarize clips. We defined a set of 79 features related to students' experimentation [4], [7]. Features include basic aggregations of actions and domain-specific counts indicative of skill demonstration. Examples include: number of trials run, counts related collecting unconfounded data, a count for changing the variable stated in a hypothesis, and number of simulation pauses. With clip labels and summary features defined, we next describe the model building process.

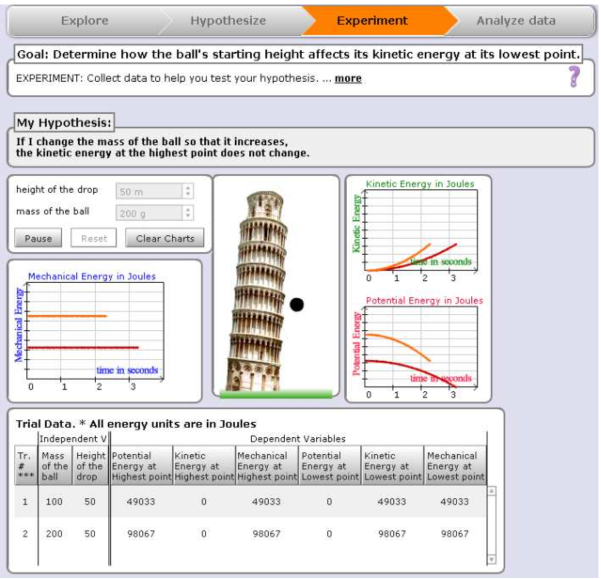
2.3 Extracting Models from the Data

To construct our models, we employed an approach that mixed traditional data mining, iterative search, and domain expertise, discussed in [7]. This procedure yielded two models, one for each skill, that take as input a clip (note that each student contributes multiple clips) and, by examining the clip's feature values, predict if a student demonstrated skill in that clip. Briefly, this procedure worked as follows.

First, human coders tagged three sets of clips, a *training set*, a *validation set*, and a *held-out test set* (more detailed descriptions of these sets appears in [7]). Next, an initial *candidate feature set* of promising predictors was selected by finding features that correlated moderately with the skill labels from the training set. This reduced the original 79 features to 11 candidate features. Then, a manual backwards elimination search was performed to find the optimum feature set yielding the best predicting model. At each stage of the search, a feature with low theoretical support was removed from the candidate feature set. Theoretical support for a feature was determined by a domain expert based on theory on features indicative (or not indicative) of skill demonstration. Then, J48 decision trees were built from the candidate feature set and training clips to yield a *candidate model*. The candidate model was then tested against the *validation set* of clips to determine how well it predicted and kept if it predicted better than its predecessor. This process repeated until predictive performance no longer improved for the validation clips. The rationale for using decision trees is described in [4].

Predictive performance was measured using A' [10] and Cohen's Kappa (κ). A' is the probability that when given two clips, one labeled as demonstrating skill and one not, the model will correctly identify which clip is which. A model with A' of 0.5 performs at chance, 1.0 indicates perfect performance. κ measures how well the model agrees with the human label; 0.0 indicates chance-level performance, 1.0 indicates perfect performance.

Once optimal models were found, their final predictive performance was measured against the *held-out test set* containing clips not used in model construction. This step was needed to better estimate true model goodness since the validation set was used during model selection. In prior versions of these models [4], cross-validation was conducted at the student-level, where students were included in either the training or test folds, validating that the models could generalize to new students.



Goal: Determine how the ball's starting height affects its kinetic energy at its lowest point.

EXPERIMENT: Collect data to help you test your hypothesis. ... more

My Hypothesis:

If I change the mass of the ball so that it increases,
the kinetic energy at the highest point does not change.

height of the drop: 50 m

mass of the ball: 200 g

Trial Data. * All energy units are in Joules

	Independent V		Dependent Variables					
Tr. # ***	Mass of the ball	Height of the drop	Potential Energy at Highest point	Kinetic Energy at Highest point	Mechanical Energy at Highest point	Potential Energy at Lowest point	Kinetic Energy at Lowest point	Mechanical Energy at Lowest point
1	100	50	49033	0	49033	0	49033	49033
2	200	50	98067	0	98067	0	98067	98067

Figure 1. Example Free Fall Activity

2.4 Prior Results and Next Steps

In [7], we found that both assessment models could predict skill demonstration for unseen clips quite well. The designing controlled experiments model was quite good at distinguishing a clip in which the skill was demonstrated from a clip in which skill was not (A' = .94). It also matched the human coder's labels better than chance (κ = .45). Similarly, the testing stated hypotheses model performed very well, A' = .91, κ = .70. These findings meant that the data-mined models could adequately identify skill demonstration in the Phase Change inquiry activities for this sample of students.

Though their performance within the Phase Change microworld is encouraging, these metrics do not provide a measure of their generalizability to other science topics, because the models were built solely from data for the Phase Change activities. Furthermore, the model construction procedure in [7] used the same students in the training/validation clip sets as in the test set. Thus, we aim here to explore the generalizability of these models to a new Physical Science topic, Free Fall. To do so, we collected data from new students who conducted inquiry Free Fall activities, tagged their resulting clips with skills, and re-measured the models' predictive performance. Using data from a different science topic enables us to assess model transfer to different topics [cf. 1]. Using new students also enables us to assess how well these models will work for a broader range of students.

3. FREE FALL INQUIRY ACTIVITIES

The Free Fall activities in Inq-ITS (Figure 1) aim to foster understanding about factors that influence the kinetic, potential and mechanical energy of a ball when it is dropped. The two factors students could change were the ball's mass and starting height. The look-and-feel and structure of these activities is generally similar to Phase Change, but with some notable differences. First, the layout of components on the screen was redesigned to improve the organization of information and instructions. Second, the number of factors the student could manipulate was smaller (2 here versus 4 in Phase Change). Third, students could only specify one hypothesis in total. Fourth, students were shown three graphs in the "experiment" phase to track each of the dependent measures over time. Finally, unlike Phase Change, the table showing the results of students' trials was always visible. Next, we describe which clips were tagged to test the generalizability of the models.

4. DATA SETS

We collected data from 292 eighth grade students' interactions with the Free Fall activities. None of these students were part of the original data set used to construct the models. Students attended three different schools in suburban Central Massachusetts. All had prior experience conducting inquiry in Inq-ITS for topics other than Free Fall. Students engaged in at most five different Free Fall activities. As per the text replay tagging process, clips were distilled to cull out student actions relevant to hypothesizing and collecting data. In total, 1580 clips were generated.

Since tagging all clips would be time consuming, we selected a representative set of clips. One approach for selecting clips for the test set is to apply student-level stratification when choosing clips to code, so that each student is equally represented in the data set. We note that this is distinct from student-level cross-validation, where students are distributed to either training or test folds, e.g. [6], [1]. Equally representing all students in a test set, and using students different than those used for model construction provides more assurance that such models will work for new students. In our work, this stratification was performed as follows:

- *Student-stratified test set (291 clips)*: One clip per student was randomly selected and tagged by a human coder. Only clips in which a student ran the simulation at least once were considered for selection. One student did not appear in this set, because they had no clips with at least one run. In this set, 90.0% of the clips and 87.6% of the clips were tagged as designing controlled experiments and testing stated hypotheses, respectively.

During the clip selection process, we noticed that a disproportionate number of clips had exactly 3 simulation runs. As shown in Table 1, 70.4% of all clips distilled had 3 simulation runs, and 74.6% in the student-level test set. Though these percentages may reflect actual student behavior, it is possible that some aspects of the models' performance may not be captured by stratifying solely in terms of the student. In particular, the models' performance may be impacted by different numbers of simulation runs. This is important because we aspire to have models that work for varying numbers of simulation runs, particularly since we activate scaffolding in the live system after students run the simulation [7]. To address this, we constructed a second test set that ensures clips with a given number of simulation runs are equally represented. This stratification is described below:

- *Run-stratified test set (245 clips)*: To generate a test set that balances the number of runs per clip, we determined an optimal number of clips to have per stratum. Given the distribution in Table 1, we used runs = 5, 49 clips, as the base. We then randomly select 49 clips for each stratum with exactly 2 simulation runs, 3 runs, etc. The final stratum was for clips with more than 5 runs. As in [7], we do not consider clips with fewer than 2 simulation runs, because demonstration of skill requires at least two trials to be collected. In this set, 93.1% of the clips and 83.3% of the clips were tagged as designing controlled experiments and testing stated hypotheses, respectively. Students' work could be represented more than once in this test set.

We note it would be more optimal to stratify over both runs and students, but too few clips would have been available for testing. In the next section, we present our models' predictive performance against these two held-out test sets.

Table 1. Counts of Clips by Number of Simulation Runs

Simulation Runs	# Clips Distilled in Total	# Clips in Student Strat. Test Set	# Clips in Run Strat. Test Set
< 2	167	20	0
2	91	18	49
3	1112	217	49
4	102	15	49
5	49	10	49
> 5	59	11	49
Total:	1580	291	245

5. RESULTS

We estimate how well the two inquiry skill assessment models built for one science topic, Phase Change, can predict skill demonstration for another topic, Free Fall, and a new sample of students. Generalizability is estimated by measuring how well the models predict skill demonstration in two held-out test sets containing clips pertaining to Free Fall activities. In the first test set, clips were randomly chosen via student-level stratification. Given our interest in understanding how well the models work at finer grain-sizes [7] and the earlier finding that clips with exactly 3 simulation runs were over-represented, we constructed a second test set. This set had clips randomly chosen to ensure a balanced number of clips with a given number of simulation runs. Performance is again measured using A' and Kappa (κ), though we also report precision and recall for our models for completeness. We focus on these metrics because they compensate for successful classification by chance, which is important given the imbalance in clip labels. Furthermore, as will be shown below, most of the models' precision and recall values are near maximum, whereas the A' and κ are more varied. Thus, we believe A' and κ may better reflect models' limitations.

5.1 Student Stratification Performance

As shown in Table 2, both models performed quite well at predicting clips in the student stratified test set. The designing controlled experiments model could distinguish a clip in which skill was demonstrated from a clip which it was not at a rate of A' = 90%. It also highly agreed with the human coder's ratings, κ = .65. Performance for the testing stated hypotheses model was also high, A' = .91, κ = .62. These findings imply that the detectors built for Phase Change generalize to another Physical Science topic, Free Fall, and to an entirely new student sample, under student-level stratification.

Recall this set has exactly one randomly chosen clip per student. Furthermore, as shown in Table 1, a majority of these clips had exactly 3 runs. Though a majority of students may run exactly three trials, providing credence to being able to use the detectors as-is to assess students, the models' performance may differ based on the number of trials collected. We turn next to performance on the run-level stratification test set.

5.2 Run Stratification Performance

As shown in Table 3, the performance profile on the run stratified test set was different than on the student stratified test set. Though the performance of the testing stated hypotheses model remained high (A'=.78, κ=.59), performance dropped for the designing controlled experiments model, particularly for raw agreement with labels (e.g. κ) (A' = .84, κ = .26). We inspected these results more closely by recalculating the metrics for each stratum of 49 clips. As shown in the bottom of Table 3, when model confidence is not taken into account (κ), the designing controlled experiments model had very low agreement with human labels for all run-levels (κ = .08 - .17) with the exception of clips with exactly 3 simulation runs (κ = 1.00). The testing stated hypotheses model fared better on agreeing with human labels on all strata (κ = .40 - .78) except for clips with exactly 4 simulation runs (κ = .00). When model confidence is taken into account (A'), both models could distinguish clips that demonstrated skill from those that did not fairly well on each strata, with the exception of the designing controlled experiments for at least 5 simulation runs (A' >= .61).

Table 2. Overall Performance on Student-Stratified Test Set

	Designing Controlled Experiments			Testing Stated Hypotheses	
	True N	True Y		True N	True Y
Pred N	26	20	Pred N	21	7
Pred Y	3	242	Pred Y	15	248
	Pc = .99, Rc = .92			Pc = .94, Rc = .97	
	A' = .90, K = .65			A' = .91, K = .62	

* Pc = precision; Rc = recall

In summary, both models performed well under student-level validation. However, under run-level validation, the testing stated hypotheses model remained strong while the designing controlled experiments models' performance suffered. In the next section, we discuss the implications of these finding on generalizability.

6. DISCUSSION AND CONCLUSIONS

We investigated whether data-mined models that assess two inquiry process skills for activities in one science topic [7] could be reused as-is for assessing those same skills for a new topic and new student sample. To explore this, we collected a new set of student interactions for the topic, employed text replay tagging [9], [4] in which student interactions (clips) were labeled by humans with skill demonstration, and measured our models' ability to predict those labels. The overarching goal of this process is to measure the degree to which these models can enable scalable, reliable performance-based assessment of the inquiry skills as students conduct inquiry within simulations [8].

Central to this work was choosing the clips to code that would yield good estimates of model performance, since coding all student interactions would be too laborious. One approach was to represent the new students equally in the held-out test set. We noticed that when we stratified this way there was an imbalance in clips for an important kind of student interaction indicative of skill, the number of times students ran the simulation. As such, we constructed a different held-out test set that ensured an equal representation over the number of simulation runs.

Under student-level stratification we found that the assessment models of each skill performed quite well in this new domain and new sample of students. These findings provide evidence that the models can be applied as-is to new topics without retraining [cf. 1]. Under run-level stratification, a different performance profile for the models emerged. The testing stated hypotheses assessment model still maintained high performance providing even stronger evidence of its generalizability. However, performance for the designing controlled experiments detector decreased. This model worked best for clips with exactly three simulation runs, the most prominent kind of clip; performance on other clips was poorer. Though performance was poorer, if the distribution of clips with given numbers of runs (Table 1) is representative of the student population we aim to assess, this model still can be used to assess in the new topic. As a side note, we did examine why performance was hindered (not presented in the results section). We found that clips that were misclassified primarily fell under a branch of the decision tree with features reflecting domain complexity (the number of variables changeable by the student). One possible way to improve generalizability would be use ratio-based features (e.g. percent of pairwise controlled trials over all possible pairs of trials) instead of a raw counts [7] for handling domain complexity.

Table 3. Performance on Run-Stratified Test Set

Designing Controlled Experiments					Testing Stated Hypotheses				
	True N	True Y				True N	True Y		
Pred N	16	60			Pred N	22	5		
Pred Y	1	168			Pred Y	19	199		
	Pc = .99, Rc = .74					Pc = .91, Rc = .98			
	A' = .84, K = .26					A' = .78, K = .59			
Runs	A'	K	Pc	Rc	Runs	A'	K	Pc	Rc
2	1.00	.08	1.00	.18	2	.84	.71	.89	.94
3	1.00	1.00	1.00	1.00	3	.88	.40	.91	.98
4	$	.00	1.00	.98	4	.84	.00	.90	1.00
5	.66	.14	1.00	.66	5	.70	.51	.89	.98
>5	.61	.17	.97	.76	>5	.79	.78	.98	.98

* Pc = precision; Rc = recall

$ = A' could not be computed because only one class label was present

This paper offers two contributions towards assessing the generalizability of data-mined models used to assess students' skills. First, like prior work [1], [3] we measure the transferability of models built for one task to a new task and new set of students. In our case, we applied data mining to assess students' inquiry skills within physical science simulations. Though we have increased evidence of models' generalizability, we note that the look-and-feel and task structure of the physical science activities were generally similar. For other science domains like biology, the nature of the experimentation process may differ; further research is needed to determine if our models will generalize to entirely new types of tasks and science domains (cf. [8]).

Second, we showed how different kinds of stratification in such a test set can reveal limitations on the performance of data mined models. In particular, the ways in which a model will be used should be considered when considering generalizability. In our work, we aim for our models to be reusable to assess all students, trigger scaffolding [8], and work regardless of how much data the student collected [7]. Thus it was essential for us to consider performance in the new simulation at the run-level since this is the granularity at which we aim to assess student work and provide scaffolding. Stratifying on other variables such as the total number of student actions or the specific inquiry activity in question [cf. 1, 3] may reveal other differences in performance. Considering these additional points may provide more evidence to the reusability of data-mined models in different contexts or reveal limitations in the models that can be addressed to improve performance in specific cases.

7. ACKNOWLEDGEMENTS

This research is funded by the National Science Foundation (NSF-DRL#0733286, NSF-DRL#1008649, and NSF-DGE#0742503) and the U.S. Department of Education (R305A090170 and R305A120778). Any opinions expressed are those of the authors and do not necessarily reflect those of the funding agencies.

8. REFERENCES

[1] Baker, R.S.J.d, Corbett, A.T., Roll, I., and Koedinger, K.R. Developing a Generalizable Detector of When Students Game the System. *User Modeling and User-Adapted Interaction*, 18, 3 (2008), 287-314.

[2] Blikstein, P. Using Learning Analytics to Assess Students' Behavior in Open-Ended Programming Tasks. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (Banff, AB, CA 2011), 110-116.

[3] Ghazarian, A. and Noorhosseini, S. M. Automatic Detection of Users' Skill Levels Using High-Frequency User Interface Events. *User Modeling and User-Adapted Interaction*, 20, 2 (2010), 109-146.

[4] Sao Pedro, M.A., Baker, R.S.J.d., Gobert, J.D., Montalvo, O., and Nakama, A. Leveraging Machine-Learned Detectors of Systematic Inquiry Behavior to Estimate and Predict Transfer of Inquiry Skill. *User Modeling and User-Adapted Interaction* (2011), 1-39.

[5] Siemens, G. Learning Analytics: Envisioning a Research Discipline and a Domain of Practice. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (Banff, AB, CA 2012), ACM, 4-8.

[6] Pardos, Z., Baker, R., Gowda, S., and Heffernan, N. The Sum is Greater than the Parts: Ensembling Models of Student Knowledge in Educational Software. *SIGKDD Explorations*, 13, 2 (2011), 37-44.

[7] Sao Pedro, M., Baker, R., and Gobert, J. Improving Construct Validity Yields Better Models of Systematic Inquiry, Even with Less Information. In *Proceedings of the 20th Conference on User Modeling, Adaptation, and Personalization* (Montreal, QC, Canada 2012), 249-260.

[8] Gobert, J., Sao Pedro, M., Baker, R., Toto, E., and Montalvo, O. Leveraging educational data mining for real time performance assessment of scientific inquiry skills within microworlds. *Journal of Educational Data Mining*, 4, 1 (2012), 111-143.

[9] Baker, R.S.J.d., Corbett, A.T., and Wagner, A.Z. Human Classification of Low-Fidelity Replays of Student Actions. In *Proceedings of the Educational Data Mining Workshop held at the 8th International Conference on Intelligent Tutoring Systems, ITS 2006* (Jhongli, Taiwan 2006), 29-36.

[10] Hanley, J.A. and McNeil, B.J. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, 143 (1982), 29-36.