The Calculus Dashboard - leveraging intelligent tutor techniques to provide automated fine-grained student assessment

Kristian Kime
Computer Science Department
Brandeis University
Waltham, MA 02453, USA
Email: kristian@brandeis.edu

Timothy Hickey
Computer Science Department
Brandeis University
Waltham, MA 02453, USA
Email: tjhickey@brandeis.edu

Rebecca Torrey
Mathematics Department
Brandeis University
Waltham, MA 02453, USA
Email: rtorrey@brandeis.edu

***Abstract*—Automatic grading systems, such as WebWork, are becoming much more widely used as they relieve the instructor from needing to grade student work, provide students with automatic feedback, and can allow for immediate resubmission. They have also been shown to improve the effectiveness of teaching and learning. In this paper, we apply Item Response Theory (IRT) to a large WebWork Calculus homework dataset to provide a skill level for each student and item characteristics curves for each problem which we then show accurately predict the probability a given student will get a particular problem correct. A student's skill level at the end of a course represents a kind of summative assessment which can be used to accurately predict how well they would answer future questions, and hence is perhaps a better indicator of subject mastery than the grade on the final exam. We also apply the Performance Factors Analysis (PFA) approach to our data and use it to provide a more fine-grained analysis of the student's mastery. PFA requires labeling each problem with a set of skills that are required to solve that problem. It produces a formula that accurately predicts the probability that students will correctly answer a new question based on their previous answers. We use the PFA approach to produce a dashboard for every student which isolates their mastery of different skills. Our dataset consisted of 703,743 attempted solutions to 243 different questions by 1609 students in 87 sections of Calculus I at a large university. Both the IRT and PFA approaches produced accurate predictions, and PFA enabled a dashboard to be constructed for each of the students.**

I. Introduction

In this paper we report on our categorization and analysis of a large Calculus homework dataset in order to develop a student dashboard framework. The main tools we use are Item Response Theory (IRT)[1][2] and Performance Factors Analysis (PFA)[3]. These are statistical tools that have been used in intelligent tutoring systems[4] [5] [6] [7] [8] but to our knowledge these approaches have not previously been applied to provide dashboards for students and teachers in conventional classes. Most learning dashboards, including conventional grade reporting interfaces in classes, rely on ad hoc calculations of ability and are not based on statistically supported approaches.

This framework is intended to take advantage of the growing number of computerized question systems in use in schools and the large amounts of data they generate. With these systems comes the chance to have a more detailed understanding of the knowledge level of students and to use that information to tailor classes and support to individuals' needs. But to do this we need a method for turning the raw information of questions and answers into something that a teacher (or learning system) can easily assess.

As a starting point we analyzed the performance of over a thousand students who had used WebWork [9] to solve Calculus problems. First we examined the data using Item Response theory (IRT). IRT is a commonly used technique that allows for the prediction of student ability and the estimation of certain item parameters for each question. These in turn can predict the probability that a student will get a given question correct. This model gives us a solid baseline for our dashboard.

Next we considered Performance Factors Analysis (PFA). PFA is a more recent approach that breaks down problems by various skills. We had a human expert perform an analysis and categorization of the questions in the dataset to assign skills to them. Using that breakdown and the student answer data, we computed parameters for those skills. Those were in turn used to predict students' outcomes. One advantage one PFA is that we can look not just at overall student ability but also at how good a student is with a particular skill.

The IRT and PFA information, when properly utilized, can provide a number of insights that can assist students and teachers and also improve our understanding of the educational process. For example a dashboard would make Just In Time teaching much more practical. A teacher would be able to see not only which students were lagging behind but also what skill areas they need to work on. It would also allow a more detailed comparison of different pedagogies by showing the difference in skill mastery between students who had been in courses with different approaches. Alternatively we would be able to look at metrics across students and see how hard different skills were overall and potentially adjust courses accordingly.

The results of our analysis provide evidence for the utility of this approach for developing a student performance dashboard

978-1-5090-5920-1/17/$31.00 ©2017 IEEE

which can provide an effective estimate of student understanding of Calculus concepts based on their performance data on homework questions.

II. The Data

The work in this paper is based on anonymized records of 1609 students in 87 sections of Calculus I at a large university from Fall 2016. This contains records of 703,743 attempts at answering 243 different questions using the WebWork system.

WebWork is an open source online homework system, primarily used for mathematics. It has a large bank of questions created and shared over the years by various instructors. The questions vary in format. For example, some questions ask students to enter a function (e.g., entering the derivative of a given function), some are multiple choice, some ask students to compute numerical values (e.g., approximating a definite integral from a table of values), etc. Figure 1 shows a sample WebWork problem. All the questions used in this study were

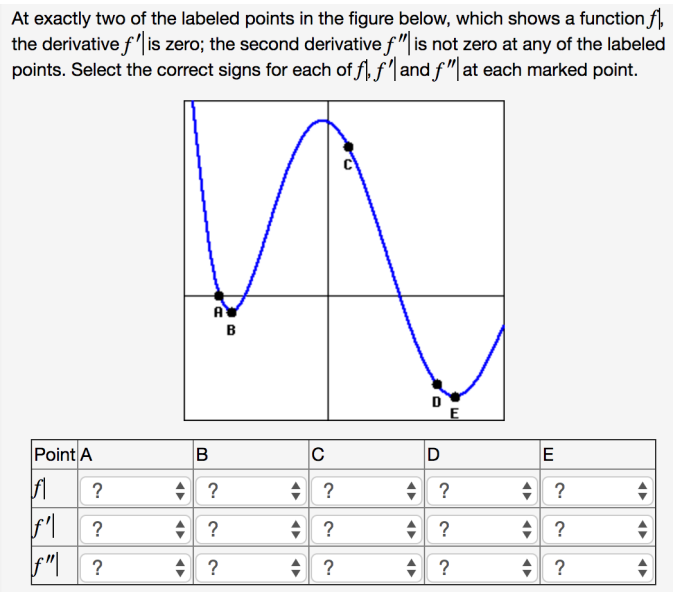

Fig. 1. WebWork Sample Problem

created by faculty at the university from which we obtained the data.

In our data set, each question attempt has the following fields:

- anon_id: a random six digit number that uniquely identifies a given student.
- problem_source: the source file for the problem, in the OPL "Library" directory of the WebWork system.
- correct_string: a string of ones and zeros indicating which part(s) of the problem are correct; thus, if there are three parts to the problem and the submission had the first and third correct, this will be 101.
- timestamp: the unix timestamp (milliseconds from Jan 1, 1970) of the submission.
- answer_string: the student's answers to the different parts of the problem, tab separated.

Students received immediate feedback on the correctness of their solutions when using WebWork, but they were limited to making at most 6 attempts at any one problem. Each attempt was recorded in this dataset.

In order to simplify our efforts we choose to take a strict definition for correct. An answer was only considered correct if all of the parts of that question were answered correctly on their first attempt. There is some evidence which suggests that the additional work to handle partial credit "does not appear to be justified" [10]. Even with this choice there is still the question of how many attempts a student is allowed. Again, to simplify our analysis we considered only the first solution submitted for any problem by a student and ignored their later attempts on that problem. Incorporating partially correct answers and multiple attempts is a subject for future research.

III. Prediction

The first research question we want to ask is

- How can we use the data to predict how well a student will do in answering a particular Calculus question?

We are looking for an approach that will allow us to predict a student's likelihood of correctly answering a specific question based on the history of their answers to all previous questions in the dataset and possibly a record of the performance of other students on that question.

Thus the result we want is a table which provides a "correctness probability" for each student and each question (and perhaps also each attempt at that question), where the probability is computed using only the student's responses for previous questions. We may however possibly allow all the responses of all other students for all questions, since we could assume this to be an accurate representation of a well-known population of students.

IV. Item Response Theory

In this section we describe how we use Item Response Theory to create a model of the data set which allows for very accurate prediction of student correctness on particular problems.

Since its conception in the 1950s and 60s by Lord [1] and Rasch[2] Item Response Theory (IRT) has become a very popular method for studying how students perform on tests. It is currently used with large scale tests like the SATs and the GRE [11]. IRT has several variations but for our purposes they are all generally concerned with estimating a skill level for each student (θ) and several parameters for each question. The questions parameters can then used in a function $p_i(\theta)$ that predicts the probability a student with skill θ will get question i correct. θ can be thought of as the number of standard deviations away form the norm a student is and is usually a value between -4 and 4.

The variation we use in this paper is the two parameter model, meaning there are two parameters for each question. The first parameter, b_i, is the "difficulty" of the question and the second, a_i, is the "discrimination". We give the precise

mathematical formulation shortly, but loosely speaking harder questions have a higher b_i. Discrimination is slightly harder to characterize. A high discrimination denotes a sharp "cliff" where students below a certain skill level have a very low chance of getting the questions right and those above have a high level. Whereas a low discrimination means that the probability of getting the question right gradually increases with skill.

The details of IRT are very technical but for reference we briefly describe the three main assumptions. First there is a unidimensional skill (θ) that captures how well the student can answer all the questions. Second that the items have "Local independence", meaning (1) questions don't have a tight relation to each other (ie there are no duplicates or very similar questions etc) and (2) students are answering questions independently (no group work or cheating). Third the students chance of getting any question right can be modeled by an a mathematical "item response function" for the question that depends only on their latent ability θ.

Given these assumptions two parameter IRT models p_i with the following function:

$$p_i(\theta) = \frac{exp(a_i(\theta - b_i)}{1 + exp(a_i(\theta - b_i))} = \frac{1}{1 + exp(-a_i(\theta - b_i))}$$

Note this function has the nice property that if θ and b_i are equal, the student has a 50% chance of getting the question right, i.e. when the students skill level equals the problem difficulty they have a fifty-fifty chance of answering correctly.

To help visualize this function we present an example. Fig. 2 shows the probability function for the sample problem shown in Fig. 1 for all θ from -4 to 4. Here we see a student who is one standard deviation below the mean ability will have about a 30% probability of getting that question correct on the first try, while a student who is one standard deviation above the mean in ability has about a 60% chance of answering it correctly. Note, we only count an answer to that question as correct if all fifteen of the subproblems (the signs of the function and its first and second derivatives at 5 points) are answered correctly on the first of up to six attempts. This is clearly a fairly hard problem for beginning Calculus students.

The details of the computation of θ, b_i and a_i are beyond the scope of this paper but there are several key features we take advantage of that are important to note. First while all the parameters must be initially estimated for a particular set of students and responses together, once the b_is and a_is are computed, we can use those values on out of sample students to predict their θs. Second a student doesn't need to have answered all the questions to get a θ value.

This means we can use some of our data as "training" data to figure out the question parameters and then use those to figure out a students skill level at various points in the semester. This in turn allows us to predict how a student will do on future problems. For example we can look at a student after they have answered the first i questions, find their current θ, and then use that to predict their probability of getting question $(i + 1)$ correct.

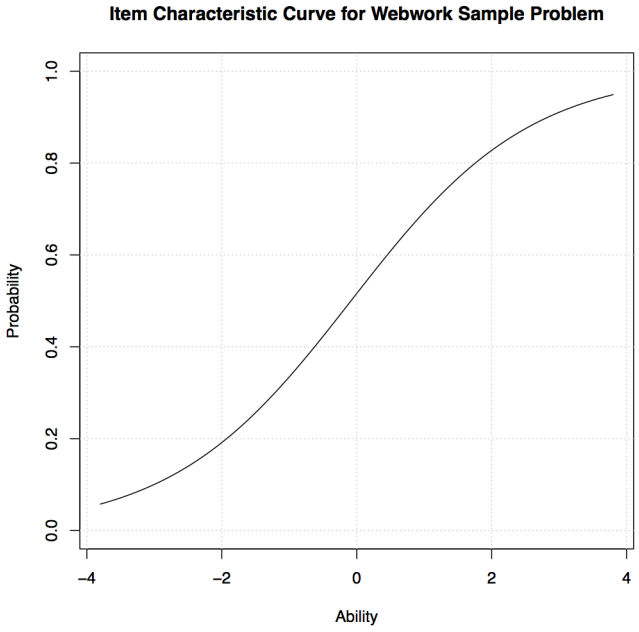

Fig. 2. IRT Probability function with difficulty b=0.064 and discrimination a=0.752

V. Applying IRT to the Data

To test the effectiveness of IRT in predicting student performance we picked 80% of the students (1287 of 1609) at random and used their data as a "training" set to calculate the parameters for all 243 problems. This left us the data for the remaining 20% of the students (322) to use in "testing" to see how well IRT could predict their answers. Fig. 3 shows the probability curves that were computed for all of the problems.

We see that two of the curves have negative slope meaning that the higher the skill level of the student the less likely they are to answer the question correctly. This seems paradoxical at first, but these are examples of difficult problems with simple guessable answers. For example, one problem is to take the derivative of

$$\arctan(x) + \arctan(1/x)$$

The answer happens to be zero as lines of slope x and $-1/x$ are perpendicular so these two arctangents sum to $\pi/2$. Many low skill students might miss that geometric interpretation or incorrectly compute the derivative algebraically, but could still simply adopt a strategy of guessing zero for problems that seem too challenging, just so they get credit for at least trying the problem.

To analyze the accuracy of the IRT prediction for the 322 students in the testing set we used their skill levels and the problem difficulties to predict their likelihood of answering each of the 243 problems. As mentioned above the problem difficulties were determined by applying IRT on the training set. Then for each student in the test set we stepped through all of the questions in order. First using their answers on all of the previous questions to determine a θ and then using that

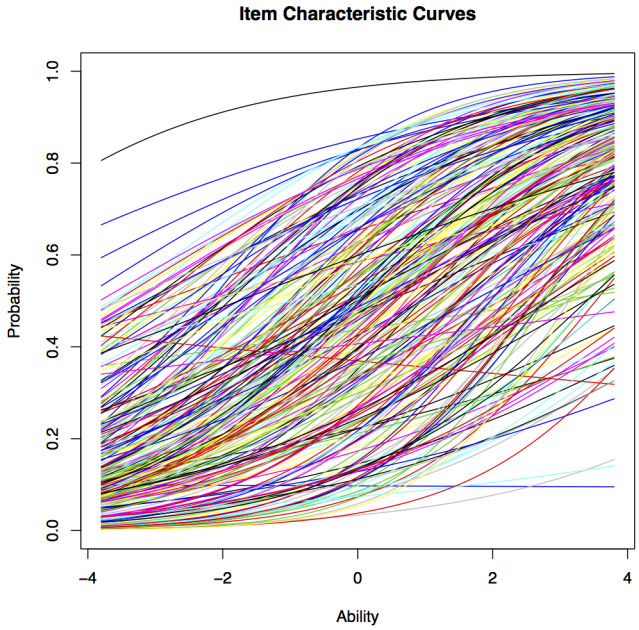

Fig. 3. Item Characteristic Curve

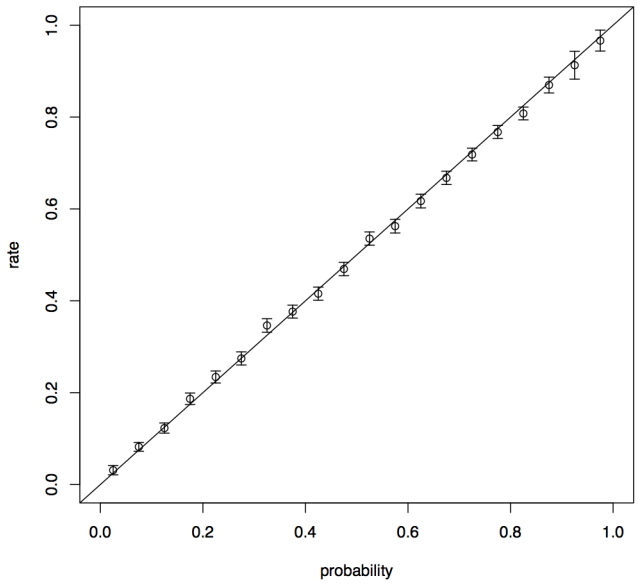

Fig. 4. IRT accuracy for binned predictions. Predictions were put into 20 bins each of which span 5 percentage points. On the X axis each "bin point" was placed at the center of the percentage range it spans (e.g. the bin [0,0.05] is plotted at 0.025). On the y axis the height of the point represents the rate at which the questions in that been were answered correctly in practice. Ideally each point would be directly on the "ideal" line.

to compute a probability for getting the next question right. In other words for each student and a given question i we used answer data from questions 1 to $(i-1)$ to compute the student's θ at that time and then used that to predict their probability to get question i correct.

This gave us 243*322=78246 predictions. We then sorted those predictions into 20 bins based on the predicted chance the student would get the question correct ([0,.05],[.05,.10],..) and calculated the percentage of those student/problem pairs in that bin that were actually correct. This allowing us to visually compare how closely the actual success rate was to what we predicted[12]. For example we would hope the actual percentage of correct answers in the 20-25 bin would be between 20% to 25%. These accuracy points are plotted in Fig. 4 and one can see that the accuracy is very high as all the points are very near to the "ideal" line overlay which represent actual accuracy exactly matching prediction.

The error bars around each point represent the 95% confidence intervals for the probabilities and they are proportional to the number of elements in the bin. We calculate the 95% confidence intervals for the proportion of correct answers in each of the bins in the standard way as

$$p \pm 1.96 * \sqrt{p(1-p)/N}$$

where p is the proportion of correct answers in the bin and N is the total number of answers in the bin. As N gets larger, the confidence interval becomes smaller.

This demonstrates both that there is a wide range of abilities among the students in the sample and a wide range of difficulties among the problems, and that even with this variability, the probability that a student will correctly answer a question can be estimated with a great degree of accuracy simply using their estimated ability θ and the problem difficulty function P_i.

While a strong case can be made that mastery of calculus could be better represented by a multidimensional estimate of ability (and indeed that is the premise of the Performance Factors Analysis approach we discuss in the next section). Our analysis though shows that the IRT approach does provide a very accurate estimate of the probability that students will answer any particular question correctly.

VI. Performance Factors Analysis

Our next goal is to see whether we can create a skills dashboard and use that to make decent predictions of the probability a student will correctly answer a question based on the skills required for that question and the student's previous experience on questions involving those skills.

To this end we rely on the Performance Factors Analysis (PFA) approach. PFA grew out of the desire to take Knowledge Tracing systems like Learning Factors Analysis [13] and extend them so they no longer require that there be a one-to-one correspondence between skills and questions. Thus in the PFA approach a student's level of mastery for a particular skill can be estimated by their performance on a set of problems provided the multiple skills required for each problem are known.

PFA [3] builds a model of student performance based on the student's successes and failures on a set of questions. Each

question is assumed to be tagged with a set of skills needed to correctly answer that question. The PFA will generate a probability that a student will successfully answer a question based solely on the tagged skills for that question and the student's history of success and failure for each of those skills.

For example, let A_k be the set of skills needed to answer a certain question Q_k and s_{ij} and f_{ij} be the number of times that the student i has succeeded (s_{ij}) and failed (f_{ij}) in answering questions requiring skill $a_j \in A_k$. The PFA model is a logistic regression which finds the best fit parameters $\beta_j, \gamma_j, \rho_j$ such that the probability that student i will correctly answer question Q_k requiring the skills A_k is given by

$$p_{ik}(m) = \frac{1}{1 + e^{-m_{ik}}}$$

where $m_{ik} = m(i, A_k)$ is given by

$$m(i, A_k) = \sum_{j \in A_k} \phi_{ij}$$

where ϕ_{ij} is the success factor for student i and skill j

$$\phi_{ij} = \beta_j + \gamma_j s_{ij} + \rho_j f_{ij}.$$

This has the nice property that it always produces a value between 0 (when $m \to -\infty$) and 1 (when $m \to \infty$). We typically have $\gamma_j \geq 0$ and $\rho_j \leq 0$, and as is usually done in PFA analyses, we enforced the constraint that $\gamma_j \geq 0$. In this case, if two students have the same history of success and failure for all of the skills except that student 1 has one more success or one fewer failure, then student 1 will have a higher probability of success than student 2.

This model assumes that the skill tagging is sufficiently precise so that mastery of those skills is all that one needs to correctly solve the problem. So two problems that require the same skills must have the same level of difficulty in this model.

One of the most important features of PFA is that it provides a probability that the student will correctly answer a question based solely on the skills required by that question and hence can produce correctness probabilities for any skill-tagged question whether or not it was in the training set. The quality of the PFA analysis clearly depends on the quality of the skill tagging, though.

VII. Applying PFA to the Data

In this section, we show that PFA can be used to make relatively good predictions of student performance on a problem, based on their performance on previous problems that incorporate at least one of the same skills.

Our first step was to annotate each of the 243 problems with labels selected from the list A of Calculus skills shown in Table I. The first four labels correspond to the "Rule of Four" classification of problems [14]. Mathematics educators often attempt to introduce and assess problems using each of these four modalities since they are all valuable ways of interacting with mathematical concepts and skills. The other skill categories were developed from the learning goals for the single variable calculus course taught at Brandeis University. This gave us some broad skills to start with, e.g., Limits & Continuity, Derivative: Definition & Concept, Derivative: Shortcuts, Applications, Antiderivatives and Differential Equations. Some of the skills were then further broken down to provide finer granularity, e.g., Trig, Logs, Product Rule, and Quotient Rule.

Once the 243 problems $\{P_k\}$ were labeled with the 25 "skills" in A we were able to apply the PFA analysis package using R to produce the three coefficients β_j, γ_j, and ρ_j for each of the 25 skills $a_j \in A$.

To test the effectiveness of PFA for modeling student performance, we randomly selected 80% of the students as a training set with the remaining 20% being a testing set. We computed the PFA coefficients for each problem using the training set of the students. We then used those coefficients to make predictions for the remaining 20% of the students. For each of these students and for each problem we used the previously computed PFA coefficients to predict the probability that students would get that problem correct based on their answers to the previous questions.

The results of applying the PFA to the WebWork data are shown in Table I. The β coefficients of the skills range from -0.693 for the Exponents skill to 0.866 for the Derivative Shortcuts. From these we can compute the probabilities (0.33 and 0.70 resp) that a student in the class will get a question requiring that skill correct assuming it is the first time they have seen that question. The γ coefficients show the correlation between the number of times a student has successfully answered a question requiring that skill and the probability that they will get the next question with that skill correct. The ρ coefficient is the same except for incorrect answers.

The values of γ range from 0 (e.g. Chain Rule) to 0.093 (e.g. Algebraic). When a skill has a γ value of zero, it indicates that getting correct answers on a problem requiring that skill does not increase the likelihood that students will get a correct answer on future problems requiring that skill. The higher the γ value, the more students seem to learn from correct answers.

The values of ρ range from -0.2 (Implicit Differentiation) to 0.014 (Product Rule). Usually we would expect ρ to have a negative value as it provides evidence that the student has not mastered the skill, but it is conceivable that even if the student gets the answer wrong, this provides important information which helps them learn the concept and hence they would have a higher probability of getting the correct answer to a similar question in the future. This is especially true for the WebWork dataset as the students have multiple attempts to answer it correctly and to learn from their mistakes. In our analysis of the PFA data though, we consider only their first answer, just as we did with the IRT analysis. Also, the students will be learning outside of the WebWork system in between problem sets (e.g., attending class, doing homework, working with friends or a tutor, etc.).

Fig. 5 shows the accuracy of the PFA prediction using the same binning approach used for the IRT prediction graph in Fig. 4. We see that the prediction is fairly accurate except in

TABLE I
PFA COEFFICIENTS

Skill	β	γ	ρ
Graphing	0.182	0.003	-0.034
Numerical	-0.297	0.025	-0.021
Verbal	-0.177	0.000	-0.147
Algebraic	-0.115	0.093	0.013
Precalc	0.487	0.013	-0.047
Trig	-0.148	0.009	-0.007
Logs	0.763	0.000	-0.104
Exponents	-0.693	0.028	0.001
Alt.Var.Names	0.196	0.000	-0.023
Abstract.Constants	0.137	0.000	-0.042
Limits...Continuity	-0.021	0.016	-0.012
Continuity..Definition	0.544	0.000	-0.183
Derivative..Definition	0.548	0.000	-0.045
Derivative..Shortcuts	0.866	0.003	-0.025
Product.Rule	-0.295	0.011	0.014
Quotient.Rule	-0.308	0.002	-0.032
Chain.Rule	-0.092	0.000	-0.006
Implicit.Differentiation	0.112	0.000	-0.200
Function.Analysis	-0.138	0.012	-0.025
Applications	-0.430	0.014	0.001
Antiderivatives	0.177	0.043	-0.023

the cases of very low and very high predictions. The large ranges there are due to the fact that PFA makes very few predictions at those extremes and we are using confidence intervals which naturally have a large range when there is a small sample size.

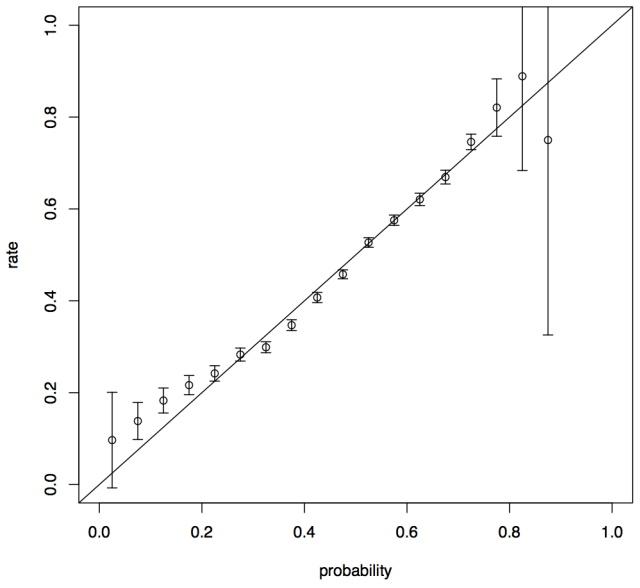

Fig. 5. PFA Predictions. This plot was created in the same way as for the IRT predictions in 4.

VIII. The PFA Dashboard

To obtain a dashboard for a student, we calculate a value for each skill representing the probability that the student would correctly answer a problem containing only that skill. The natural approach is to look at all questions the student has answered that require that skill, calculate the number of correct and incorrect answers to those questions, and then use that information and the PFA data to make a prediction for the probability that the student would correctly answer a question that only required that skill.

In Table II, we show the summary statistics for the Dashboard values for all 322 students in the test set at the end of the semester. The ranges of the dashboard values in Table II vary from a min of 0.15 for graphing to a max of 0.73 for Anti-derivatives. We might have expected higher values for the range maxima, but there are several reasons why the maximum skill levels don't reach 1.0. For example, these are low-stakes homework grades with multiple part questions and we are requiring them to get all parts correct on the first attempt, even though they are allowed six attempts. This could serve to lower the expected Dashboard values. Also, each of the problems tends to require several skills and the PFA algorithm calculates the probability of correctness for a student i on a problem k, in part, by adding up student i's success factors ϕ_{ij} for each of the skills a_j in problem k. In our dashboard, we assume that only that one skill is required which will necessarily result in a lower value for the sum. The important information provided by the dashboard is the relative rank of the student's skill level compared to all other students in the class.

Table III shows the Dashboard values for 3 randomly selected students and Fig. 6 shows a graphical version of the dashboard for the first student in Table III where we use a box and whisker plot to show the distribution of that dashboard value over all of the students in the Testing sample and we use an asterisk for the student's particular Dashboard value. This data could be used to help students see their relative strengths in the class. For example, using this combined with Table II we see that student 1 is below the 25th percentile (q1) for the Graphing skill, but at the 75th percentile (q3) for the Algebraic skill, and hence they don't need to worry as much about their Algebraic skills but should probably work on developing their Graphing skills.

Fig. 7 shows a scatter plot of the Graphing skill level versus the Numerical skill level. One thing to note is that there is a bigger range of skill for graphing problems (0.15-0.54) than for numerical skills (0.33-0.50). Moreover, there is no strong positive or negative correlation between these two dashboard items which suggests that they are relatively independent skills.

IX. Conclusions and Future Work

We have shown that the IRT and PFA tools can be used to make accurate predictions of whether students will correctly answer a particular Calculus question, based on their answers to previous questions using the WebWork application. We have further shown the PFA analysis can easily be used to produce a dashboard which gives skill levels for each student on each skill.

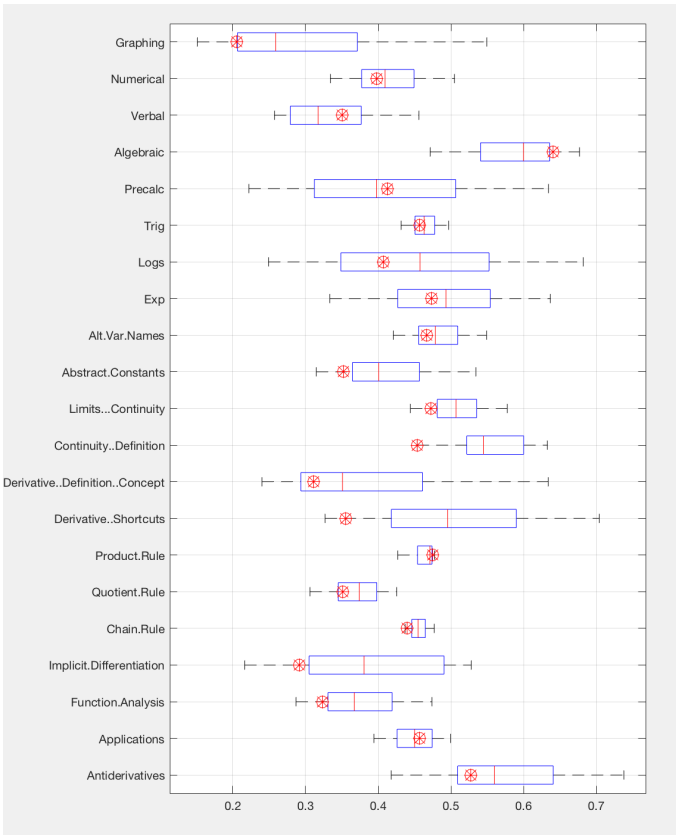

Fig. 6. Graphical Version of the PFA Dashboard for Student 1

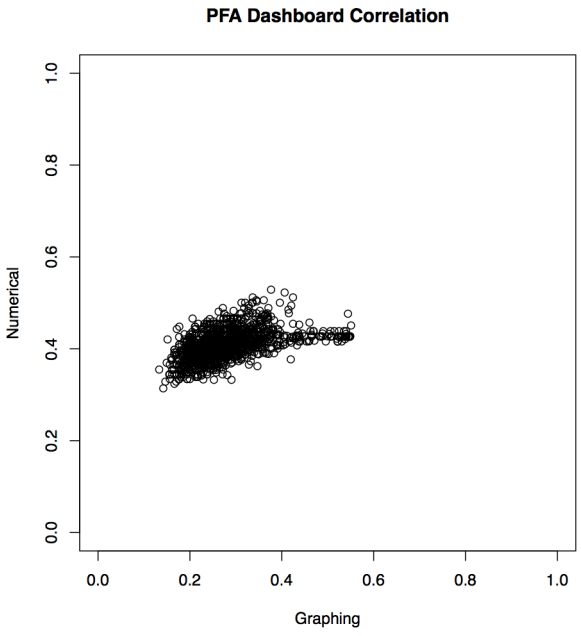

Fig. 7. PFA Dashboards ScatterPlot: Graphing vs Algebraic

TABLE II
DISTRIBUTION OF PFA DASHBOARD VALUES

skill	min	q1	med	q3	max
Graphing	0.15	0.22	0.25	0.31	0.54
Numerical	0.33	0.39	0.40	0.43	0.50
Verbal	0.25	0.28	0.31	0.35	0.45
Algebraic	0.47	0.56	0.59	0.62	0.67
Precalc	0.22	0.34	0.39	0.46	0.63
Trig	0.43	0.45	0.46	0.47	0.49
Logs	0.24	0.38	0.45	0.50	0.68
Exp	0.33	0.45	0.49	0.52	0.63
Alt.VaNames	0.42	0.46	0.47	0.49	0.54
Abstract.C	0.31	0.38	0.40	0.43	0.53
Limits.C	0.44	0.49	0.50	0.52	0.57
Continuity.Def	0.45	0.54	0.54	0.58	0.63
Derivatie.Def	0.24	0.31	0.35	0.40	0.63
Derivative.S	0.32	0.44	0.49	0.55	0.70
Product.Rule	0.42	0.46	0.47	0.47	0.47
Quotient.Rule	0.30	0.35	0.37	0.38	0.42
Chain.Rule	0.43	0.44	0.45	0.46	0.47
Implicit.Dif	0.21	0.33	0.38	0.47	0.52
Function.A	0.28	0.34	0.36	0.40	0.47
Applications	0.39	0.43	0.45	0.46	0.49
Antiderivatives	0.41	0.53	0.56	0.60	0.73

TABLE III
SAMPLE STUDENT DASHBOARDS

skill	student 1	student 2	student 3
Graphing	0.20	0.27	0.21
Numerical	0.39	0.43	0.38
Verbal	0.35	0.35	0.38
Algebraic	0.64	0.60	0.62
Precalc	0.41	0.49	0.41
Trig	0.45	0.46	0.45
Logs	0.40	0.53	0.38
Exp	0.47	0.54	0.49
Alt.Va	0.46	0.49	0.44
Abstract.C	0.35	0.43	0.40
Limits.C	0.47	0.52	0.47
Continuity.D	0.45	0.54	0.54
Derivative.D	0.31	0.35	0.34
Derivative.S	0.35	0.53	0.40
Product.Rule	0.47	0.46	0.47
Quotient.Rule	0.35	0.38	0.32
Chain.Rule	0.44	0.45	0.44
Implicit.Dif	0.29	0.38	0.38
Function.A	0.32	0.36	0.32
Applications	0.45	0.46	0.45
Antiderivatives	0.52	0.60	0.54

The results we have presented suggest that it would be interesting to modify WebWork or a similar application so that it could provide a dashboard for every student and a summary dashboard for the instructor. This modification would require that each problem be tagged with a set of skills and that the PFA analysis should be recomputed after each homework assignment to recalculate skill levels.

The main benefit of the PFA approach is that it disentangles the skills from the problems and provides an isolated measure for that particular skill, even though in our case almost every problem requires multiple skills. These skill levels, in turn, can be used to provide an accurate estimate of the probability the student will correctly answer a new question, solely in terms of the skills required for that question.

One feature of the data which we have not considered in this analysis is that students were given immediate feedback on their submitted answers and were allowed to resubmit up to 6 times. It would be interesting to explore how to include the performance on resubmission data into the dashboard view so as to provide more information on their mastery of the material.

Another feature that has not been incorporated is that many of the problems were multi-part and our system only counted their answer as correct if all of the parts were correct, otherwise it was counted as completely wrong. It would be interesting to allow for non-binary measures of correctness, or alternately to categorize the skills needed for each of the subparts and use that information in the PFA analysis.

These PFA-based dashboards could have many benefits if built into a Problem Solving Learning Environment like WebWork. For example, they could help students decide which areas to focus on if they see their understanding is below the class average. It could also be used by teachers to determine which topics need more coverage in class and to assign personalized homework for students based on their dashboard values. It might also be used to design questions that would more effectively gauge the student's mastery of a particular skill, so as to improve the dashboard quality and the concomitant confidence in the students' mastery.

If this line of research continues to provide promising results, it would be possible to augment course grades for STEM classes with an automatically computed PFA Skill Report. This could clearly help students develop a more nuanced understanding of their mastery of a certain subject. Rather than thinking, "I'm bad at Calculus", they could see that there are some topics they have mastered and others that need more work. This kind of formative assessment might even be superior to the usual summative assessments (e.g., a final exam) and could be used for a more fine-grained notion of prerequisite requirements for future classes.

REFERENCES

[1] F. M. Lord, "The Relation of Test Score to the Trait Underlying the Test," *ETS Research Bulletin Series*, vol. 1952, no. 2, pp. 517–549, Dec. 1952.

[2] G. Rasch, *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Oxford, England: Nielsen & Lydiche, 1960, vol. xiii.

[3] P. I. Pavlik, H. Cen, and K. R. Koedinger, "Performance factors analysis –a new alternative to knowledge tracing," in *Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling*. Amsterdam, The Netherlands, The Netherlands: IOS Press, 2009, pp. 531–538.

[4] M. C. Desmarais and R. S. Baker, "A Review of Recent Advances in Learner and Skill Modeling in Intelligent Learning Environments," *User Modeling and User-Adapted Interaction*, vol. 22, no. 1-2, pp. 9–38, Apr. 2012.

[5] Y. Gong, J. Beck, and N. Heffernan, "Comparing knowledge tracing and performance factor analysis by using multiple model fitting procedures," in *Intelligent tutoring systems*. Springer, 2010, pp. 35–44.

[6] J. Johns, S. Mahadevan, and B. Woolf, "Estimating Student Proficiency Using an Item Response Theory Model," in *Intelligent Tutoring Systems*. Springer, Berlin, Heidelberg, Jun. 2006, pp. 473–480.

[7] H. Mohamed, T. Bensebaa, and P. Trigano, "Developing adaptive intelligent tutoring system based on item response theory and metrics," *International Journal of Advanced Science and Technology*, vol. 43, pp. 1–14, 2012.

[8] M. Feng, J. Beck, N. Heffernan, and K. Koedinger, "Can an Intelligent Tutoring System Predict Math Proficiency as Well as a Standardized Test?" in *Educational Data Mining 2008*, 2008.

[9] M. Gage, A. Pizer, and V. Roth, "WeBWorK: Generating, delivering, and checking math homework via the Internet." in *Proceedings of the 2nd International Conference on the Teaching of Mathematics*, Crete, Greece, 2002.

[10] G. Kortemeyer, "Extending item response theory to online homework," *Physical Review Special Topics-Physics Education Research*, vol. 10, no. 1, p. 010118.

[11] "Item response theory," Apr. 2017, page Version ID: 774931612. [Online]. Available: https://en.wikipedia.org/wiki/Item_response_theory

[12] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan, "Assessing the performance of prediction models: a framework for some traditional and novel measures," *Epidemiology (Cambridge, Mass.)*, vol. 21, no. 1, pp. 128–138, Jan. 2010.

[13] H. Cen, K. Koedinger, and B. Junker, "Learning factors analysis a general method for cognitive model evaluation and improvement," in *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, ser. ITS'06. Springer-Verlag, pp. 164–175.

[14] D. Hughes-Hallett, *Applied Calculus*. John Wiley & Sons, Incorporated.