Data from a Web-based Homework Tutor can predict Student's Final Exam Score

David E. Pritchard
Department of Physics & Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139, USA.
dpritch@mit.edu

Rasil Warnakulasooriya
Department of Physics & Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139, USA.
rasil@mit.edu

Abstract: We discuss new measures in assessing the skill level of students based on their interactions with the Socratic homework tutor, MasteringPhysics. We show that the measurement of the difficulty of problems for a given student as determined by the time to first correct response, the number of incorrect responses without advice, and hints have high reliability (96%). As an immediate application, we demonstrate how item difficulty can be used to construct an item discrimination measure that would result in predicting the final exam score with a correlation of 0.634.

Introduction

Web-based instruction offer new avenues for research in learning (Mitchell, Dipetta & Kerr 2001). In this paper we consider one such online tutor, namely MasteringPhysics. One of its advantages is the ability to assess learning and the skill of students in a particular subject through measures that are unavailable in traditional methods of instruction such as time on task, number of requested hints, and number of wrong answers given en route to the solution. We demonstrate high reliability that can be achieved using such measures, and how similar methodology may be extended in predicting the final exam scores of students.

We are motivated by the desire:

1. To identify and develop more reliable (less measurement error) measures of a student's skill – a tool that opens up a vast area of future research in educational psychology. These include, but are not limited to, studies on skill acquisition (VanLehn 1996), epistemology (Hofer & Pintrich 1997), pedagogy (Mayer 2003), and studies of student learning styles such as expert-novice differences (Chi et al. 1981, Jacobson 2001).

2. Develop measures to identify students who engage in intellectual dishonesty in web-based homework: This will help improve precision in the studies by removing such data points.

3. Predict students' performance on high stakes exams based on skill measures developed.

We first discuss the pedagogy of MasteringPhysics, then introduce the measure for "difficulty" of a problem for a student, and show its reliability using the split-half method. We then show how the difficulty of a problem can be used for item discrimination analysis and the final exam score prediction.

The Pedagogy of Mastering Physics

The studies we report here were conducted with a web-based Socratic tutor available commercially as MasteringPhysics from Addison Wesley[1]. The pedagogy of MasteringPhysics is based on mastery learning (Bloom 1981) in which the time on task is increased, and feedback supplied, for each student until over 90% of the students can solve the problem. This is the reverse of most in-school instruction where the time is fixed and only the most skillful students master the material at this level. Mastery learning is implemented within a Socratic dialogue where students are provided with hints and simpler sub-problems upon request, and are given specific criticism (feedback) when incorrect answers are proposed. The hints and sub-problems are designed to impart declarative and procedural knowledge, respectively. In addition, follow-up comments or questions are frequently given to highlight important features or implications of an answer that has just been obtained. The follow-ups are designed to foster active engagement of the student (Redish, Saul, & Steinberg 1997; Sokoloff & Thornton 1997). If the students exhaust the available hints they can request the solution to a problem or a sub-problem. The homework score is based on the percentage of correct solutions obtained less a small penalty for making wrong answers or requesting hints (this discourages guessing and encourages the students to figure it out for themselves).

The absolute difficulty algorithm

The data for the study we report here comes from a class of ~340 students taking the “Introductory Newtonian Mechanics” course at the Massachusetts Institute of Technology (MIT) in fall 2003. The students were assigned weekly homework in MasteringPhysics.

In the course of enabling over 90% of the students to solve each problem, the tutor compiles a log of the interactions with each student that contains: the time to completion of a multi-part problem (T), the number of hints requested (h), the number of incorrect answers given without any advice (ina) other than “try again,” the number of incorrect answers with advice (ia), the fraction of correct answers given on first attempt to a multi-part problem (cft), and the number of solutions requested (s). Time to completion is defined as the time interval between first opening the problem and submitting the problem without accounting for any events in between.

With ~94% of the students eventually getting the correct solution, the usual assessment criterion (“Is the answer correct?”) does not adequately assess the students. It turns out that a much better (and more reliable) assessment is available by looking at the process of solution, and in particular at those interactions that indicate that students are experiencing difficulty in obtaining the solution. Our goal was to devise a difficulty measuring algorithm for a given problem by a given student. We optimized the reliability of this algorithm using the split-half method. For this purpose we chose 64 problems on various assignments given throughout the semester and divided them into two sets of 32 (called “even” and “odd”) with one problem on a given conceptual domain in each. We compute average difficulty (D) of a given set (even or odd) for a student as a linear combination of the average values of T, ina, and h. We find that a simple difficulty algorithm such as,

$$D = T + ina + h \qquad (1)$$

gives a correlation of 0.85 between the average difficulty of even and odd problems yielding a split-half reliability of 92% using the Spearman-Brown formula. It should be noted that larger values of T, ina, and h indicates a greater difficulty for a student on a given problem. This reliability can be improved (to 93%) by considering time for the first correct response (t) instead of time to completion (T). It should be mentioned that t and T are the natural logarithms of the time to first correct response (measured in minutes) and time to completion (measured in minutes), respectively. Furthermore, we are simply finding the values of the variables for a given student and are not interested in the relative standing of the student with respect to the rest of the class for a variable of interest; hence our use of the word “absolute” difficulty.

[1] David Pritchard founded the Effective Educational Technologies which makes MasteringPhysics with his son, and their family has a controlling interest in the company. Some of the IP for this company is licensed from MIT where it was developed under his direction.

Maximizing the reliability is achieved by maximizing the correlation of the average difficulty between the even and the odd problems. We look for a weighted sum of the three variables given above in the form,

$$D = \alpha{*}t + \beta{*}ina + h, \qquad (2)$$

where, α and β are the weights to be determined and the weight of *h* is fixed at 1. It should be noted that the average difficulty of both the even and the odd sets of problems is determined by (2). Therefore, the question of maximizing the correlation is a unique problem that cannot be handled by multiple-regression since *D* is unknown and is determined by the three predictors (*t*, *ina*, and *h*). In other words, α and β has to be determined simultaneously for the average difficulty on even and odd sets. We address this problem along the lines of Breiman and Friedman (1985) by minimizing

$$\frac{\sum (D_{odd} - D_{even})^2 / N}{\sum \{D_{odd} - Avg(D_{odd})\}^2 / N} \qquad (3)$$

with respect to α and β, which corresponds to maximizing the correlation (r) between the even and odd average difficulties, and hence the reliability. In (3), the sum is over students and *N* is the number of students. This procedure yields the algorithm,

$$D = 0.025{*}t + 0.248{*}ina + h, \qquad (4)$$

with a correlation of 0.897 and a reliability of 94%. Thus, 80% of the variance (which is the correlation squared) is explained by the resulting regression line. This should be contrasted with our observation that a similar split-half study for problems in the paper-based final exam only accounts for about 40% of the variance (Pritchard & Morote 2002). Thus, the derived measure of difficulty, *D*, using MasteringPhysics data reduces the error variance by about a factor of two. The reliability can be further improved, up to 96% by removing the regression outliers (Figure 1) according to Rousseeuw and van Zomeren (1990).

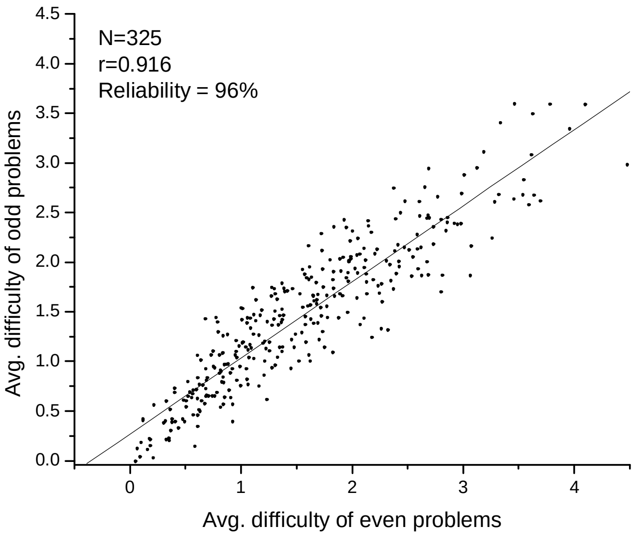

Figure 1: Correlation of average difficulty between two sets (even and odd) problems containing 32 problems each. A point corresponds to a single student. The average difficulty is calculated using 0.025**t* + 0.248**ina* + *h* and removing the regression outliers.

Final exam score prediction

We can extend the model development to predict the final exam score of students. We initially considered the variable *T* (time to completion), *h* (hints), *ia* (incorrect with advice), *ina* (incorrect not receiving advice), (*1-cft*) where *cft* is the fraction (or probability) of correct responses on the first attempt, and *s* (solutions requested) as the predictor variables of the final exam score. However, we find that (*1-cft*) is highly correlated with other predictors and leads to high (~10) variance inflation factors (VIFs). The VIFs are a measure of the inter-correlations among the predictors (Montgomery, Peck & Vining, 2001). This is understandable since not being able to respond correctly on the first attempt leads to requesting hints, and thereby taking longer time in completing a given problem. Also, ina is directly related to (*1-cft*). For these reasons we eliminated the variables (*1-cft*) and *ina* as predictors for the final. The remaining four predictors result in a correlation of 0.288 with the final exam score.

We must emphasize that the above correlation was achieved after removing the outliers as determined by using minimum generalized variance (Wilcox, 2003). For example, before removing the outliers, the correlation between the final exam score and *T* is positive; i.e. the more time students spent on the MasteringPhysics problems the better they did on the final exam. This is certainly feasible if learning is taking place in the tutor, but a closer look reveals that the reason for the positive slope of the regression line is the presence of students who were completing MasteringPhysics problems (a total of 219 problems during the semester) with an anomalously short time < 2.5 minutes (Warnakulasooriya & Pritchard, 2005), yet scoring well below the average on the final. The result is that the regression line is “dragged down” by low final scores at low values of *T*, thereby giving a positive slope. However, removing these outliers result in a negative correlation, which corresponds to the expected behavior on average.

Furthermore, we find that the fraction of problems completed in less than 2.5 minutes (*pfr*) is well correlated (r = -0.37) with the final exam score. The more problems students completed in less time the lower they scored in the final exam. This suggests either academic dishonesty and/or disinterest in the course. Such students were again removed as identified by using minimum generalized variance. The resulting model having the predictors *T, h, ia, s*, and *pfr* correlate with the final exam score at 0.514. The resulting algorithm where the final is predicted on a scale from 0 to 1 is,

Predicted final score = – 0.394**T* – 0.228**h* – 0.078**ia* – 0.062**s* – 0.472**pfr* + 0.991.

The alternating conditional expectation algorithm (Breiman & Friedman, 1985) followed by Box-Cox transformations suggest (Box & Cox, 1964) that $(T)^{-2.2}$ and $(pfr)^{2.3}$ should improve the correlation (the transformations were applied only to *T* and *pfr* since they are sufficiently monotonic). This is indeed the case - the correlation is improved from 0.514 to 0.585.

We next consider a new predictor, which we will call “*disc*” based on the idea of weighting each MasteringPhysics item by its discrimination index relative to the overall final. We define *disc* as,

$$disc \equiv \frac{\sum (D_{in} \Lambda_n{}^a)}{\sum (\Lambda_n{}^a)} \qquad (5)$$

where D_{in} is the difficulty of the nth problem for the ith student as given by (4), where the sum is over the items. Λ is the item discrimination index of the nth problem and *a* is an exponent. The item discrimination index (Λ) is essentially a correlation coefficient found by correlating the final exam score with the item difficulty D_{in}. Thus, if an item in MasteringPhysics discriminates in favor of the more skillful versus the non-skilled student, then Λ must be negative; i.e. the difficulty of the item should decrease for a student who can score well on the final exam. We find that out of the 219 items, 59 of them do not discriminate well in the sense that students who found the item more difficult scored more in the final than the students who found the item less difficult. These items were discarded before including *disc* in the final prediction algorithm.

Adding variable *disc* to the final prediction algorithm we find that the variable *h* has to be removed since it leads to high variance inflation. This is expected since *h* is included in the variable *disc* through *D*. Another problem encountered was the "wrong sign problem" for *ia*: that is, although *ia* is negatively correlated with the final in simple regression, it is positively correlated in multiple regression. This probably indicates a yet unknown variable that must be accounted for in the model. Thus, our final model only includes the "effective" predictors *T*, *s*, *pfr*, and *disc*. We note that *disc* further contain the predictors *t*, *ina*, and *h* in the form of *D*. The final exam prediction algorithm (on a scale from 0 to 1) is then,

Predicted final score = $0.474*T^{-2.2} - 0.037*s - 0.548*pfr^{2.3} - 0.409*disc\,|_{a=2} + 0.632$.

It is noted that the highest correlation is obtained for the exponent $a = 2$ (to the nearest integer) of the predictor *disc*. This algorithm results in a correlation of 0.634 with the final exam showing that it can account for 40% of the total variance. The standard error of the estimate is 0.161. Furthermore, the 95% confidence interval for the correlation is found by the method of bootstrap (Efron & Tibshirani, 1993; Wilcox & Muska, 2001) which gives (0.522, 0.695).

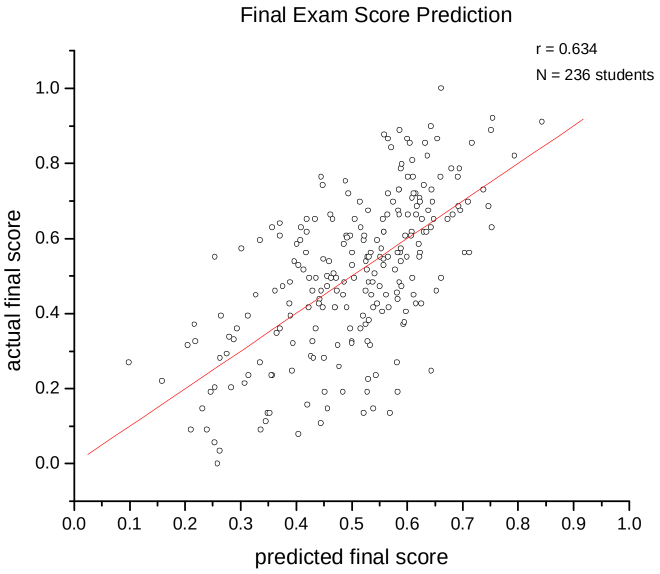

Figure 2: Predicted versus the actual final exam score using the algorithm: $0.474*T^{-2.2} - 0.037*s - 0.548*pfr^{2.3} - 0.409*disc\,|_{a=2} + 0.632$.

It may be argued that the use of the final exam score to calculate item discrimination indices (Λ) is circular since our objective itself is to predict the final exam score. However, the penalty we had to pay for using the final score as such is that we had to discard 59 problems. This may also indicate the mismatch between the final exam and the MasteringPhysics problems to some extent. We have investigated a method which would address the above objection by calculating item discrimination indices based on the average difficulty rather than the final score. This gives us an algorithm which has a correlation of 0.592 (0.485, 0.666; 95% confidence interval) with a comparable standard error of estimate, which is 0.168. The corresponding exponent (a) is found to be unity Thus, for the current data set we do not see a substantial difference between these two methods of item discrimination.

Summary and discussion

We have evidence that exceptionally high precision assessment is feasible with the data available from student interaction with MasteringPhysics, the web-based homework tutor. We have developed a difficulty measure (*D*) which yields a split-half reliability of over 92% (such reliabilities are generally considered as superior in standard educational literature). We also are capable of predicting the final exam score with a correlation over 0.5. We believe that our models are robust enough to have sufficient repeatability since we have accounted for the outliers and the variables which will lead to unreliable regression coefficients. We will report on the performance of the above algorithms for students in other classes in the future.

In the present study we have eliminated all the outliers as identified by the minimum generalized variance method. However, we may keep certain "outliers" considering a variety of other factors, which we have not done in this study. However, we do not believe that this will affect our conclusions in any drastic manner.

Also, we may be able to use a criterion other than 2.5 minutes to identify students who did not perform well in the final compared to the time they took to complete problems on average. Such a criterion might be the fraction of problems done in less than 1/3 of the average time of the problems of interest. We already have data that suggests that this may be a better identifier of such students, the results of which we will report with other improvements to the current algorithms.

Our ability to predict the final exam score using MasteringPhysics data is a measure of our ability to assess the skill of a student on an equivalent scale on the given subject matter. Since such an assessment is made over the course of the semester over several hundreds of problems with many variables that directly correlate with the student's skill, it gives us a better way to deal fairly with students' actual skill. This eliminates the high stakes nature of a final exam. Given such assessment capabilities, teachers could confidently determine students' skill without worrying about the one who miraculously passed, deserved to pass, or failed, just because of some "bad luck" (Pritchard & Morote, 2002).

References

Bloom, B. S. (1981). *All Our Children Learning.* NY: McGraw-Hill.

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations. *Journal of Royal Statistical Society*, Series B, **26**, 211-252.

Breiman, L., & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, **80**, 580-598.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science,* **5**, 121-152.

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the Bootstrap*. New York: Chapman and Hall.

Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Review of Educational Research*, **67**, 88-140.

Jacobson, M. J. (2001). Problem solving, cognition, and complex systems: Differences between experts and novices. *Complexity,* **6**, 41-49.

Mayer, R. E. (2003). *Learning and Instruction.* NJ: Pearson.

Mitchell, C., Dipetta, T., & Kerr, J. (2001). The frontier of web-based instruction. *Education and Information Technologies,* **6**, 105-121.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2001). *Introduction to Linear Regression Analysis*. NY: Wiley.

Pritchard, D. E., & Morote, E –S. (2002). Reliable Assessment with CyberTutor, a web-based homework tutor. *World Conference on E-Learning in Corporate, Government, Health, & Higher Education*, **2002**, 785-791.

Redish, E. F., Saul, J. M., & Steinberg, R. N. (1997). On the effectiveness of active- engagement microcomputer-based laboratories. *American Journal of Physics*, **65**, 45-54.

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association,* **85**, 633-639.

Sokoloff, D. R., & Thornton, R. K. (1997). Using interactive lecture demonstrations to create an active learning environment. *Physics Teacher*, **35**, 340-347.

VanLehn, K. (1996). Cognitive skill acquisition. *Annual Review of Psychology*, **47**, 513-539.

Warnakulasooriya, R., & Pritchard, D. E. (2005). Learning and problem-solving transfer between physics problems using web-based homework tutor. *EdMedia Conference Proceedings*.

Wilcox, R. (2003). *Applying Contemporary Statistical Techniques*. CA: Elsevier.

Wilcox, R., & Muska, J. (2001). Inferences about correlation when there is heteroscedasticity. *British Journal of Mathematical and Statistical Psychology*, **54**, 39-47.

Acknowledgements

This work is funded by the National Science Foundation under grant 0231268.