Enhancing the Efficiency and Reliability of Group Differentiation through Partial Credit

Yan Wang, Korinn Ostrow, Joseph Beck, Neil Heffernan
Worcester Polytechnic Institute
Worcester, MA 01609
{ywang14, ksostrow, josephbeck, nth} @wpi.edu

ABSTRACT

The focus of the learning analytics community bridges the gap between controlled educational research and data mining. Online learning platforms can be used to conduct randomized controlled trials to assist in the development of interventions that increase learning gains; datasets from such research can act as a treasure trove for inquisitive data miners. The present work employs a data mining approach on randomized controlled trial data from ASSISTments, a popular online learning platform, to assess the benefits of incorporating additional student performance data when attempting to differentiate between two user groups. Through a resampling technique, we show that partial credit, defined as an algorithmic combination of binary correctness, hint usage, and attempt count, can benefit assessment and group differentiation. Partial credit reduces sample sizes required to reliably differentiate between groups that are known to differ by 58%, and reduces sample sizes required to reliably differentiate between less distinct groups by 9%.

Categories and Subject Descriptors

K: Applications to Education. K.3: Computers and Education. I.6 Simulation and Modeling.

General Terms

Measurement, Experimentation, Reliability.

Keywords

Partial Credit, Group Differentiation, Resampling, Randomized Controlled Trial, Data Mining.

1. INTRODUCTION

The learning analytics and educational data mining communities have established a variety of well-vetted models to predict student knowledge and trace performance both within and across knowledge components (i.e., skills). The gold standard for student modeling, Knowledge Tracing (KT), has maintained its reign for almost a quarter-century despite relying on a rudimentary sequence of correct and incorrect responses to estimate the probability of student knowledge [2]. Attempts to enrich this approach have included supplemental estimates of prior knowledge to individualize predictions to each student [9], supplemental estimates of item difficulty to individualize to each problem [10], and the implementation of flexible correctness via consideration of hint usage and attempt count [12, 13, 7]. Despite these excursions, popular learning systems, including the Cognitive Tutor series, still largely rely on traditional KT to inform mastery learning [4].

In parallel, enthusiastic support has been growing for the use of randomized controlled trials embedded within online learning platforms to investigate best practices and enhance the user experience. Randomized controlled trials are the soundest approach to social science, allowing researchers to postulate causal relationships between independent and dependent variables. Within the realm of education, experimental design has historically been longitudinal, with formal pre- and post-tests, highly controlled curricula, and vast sample populations required for class-level or even school-level randomization. However, the expanding popularity of online learning platforms used for classwork and homework offers researchers an opportunity to gather data more efficiently, with fewer logistic constraints, and requiring smaller samples due to random assignment at the student-level.

The present work employs data mining methodologies on randomized controlled trial data from ASSISTments, a popular online learning platform, to assess the benefits of incorporating additional student performance data when attempting to differentiate between two user groups. The platform, created in 2002, now supports over 50,000 users around the world, providing students with immediate feedback and enhancing assessment for teachers [3]. The ASSISTments platform is an easily accessible shared tool for educational research that offers the unique opportunity to bridge the gap between the analysis of randomized controlled trials and more traditional data mining. Considering student performance variables for the purpose of group differentiation is arguably a worthy venture for both realms.

Many learning platforms assess student performance using standard binary correctness (i.e., a student's accuracy on her first solution attempt). Instead, we argue for a combination of features that better define the learning process: initial accuracy, feedback usage, and attempts required for success. The present work suggests that such features can be combined to establish a partial credit metric to enhance analytic efficiency when attempting to differentiate between two user groups (i.e., experimental conditions). It is not surprising that a more robust view of student performance can alter a researcher's ability to pinpoint the effectiveness of an intervention. Modeling numerous features per data point requires fewer data points to arrive at distinct conclusions (i.e., posttests could simultaneously be shortened and yet made more robust for both students and researchers). Previous

LAK '16, April 25-29, 2016, Edinburgh, United Kingdom
© 2016 ACM. ISBN 978-1-4503-4190-5/16/04…$15.00
DOI: http://dx.doi.org/10.1145/2883851.2883910

work has also suggested that infusing controlled assessment with learning opportunities (i.e., providing feedback or allowing multiple attempts) directly benefits robust student learning [1]. However, many researchers hesitate when considering the allowance of these features within posttests. As such, the present work seeks to validate the allowance of 'partial credit' within randomized controlled trial posttests.

Although ASSISTments employs binary scoring, feedback usage and attempts required for success can be considered in the algorithmic calculation of partial credit scores. Recent research within ASSISTments has examined the potential benefits of partial credit scoring for student modeling [7] and has validated partial credit penalizations using an extensive grid search of possible scoring procedures [6]. We extend this work by asking: Does partial credit scoring enhance the efficiency with which significant differences can be detected between groups of students within a randomized controlled trial? We define 'enhanced efficiency' as a reduction in the sample size required to reliably observe significant differences between groups (akin to enhancing power, or reducing Type II error).

2. DATASET

The dataset is comprised of log files from a previously published randomized controlled trial on the effects of interleaving skill content within a brief homework assignment [8]. The original study was conducted with a group of participating teachers from a suburban middle school in Massachusetts. Researchers worked with teachers to select content for three skills (A, B, C). A practice session comprised of twelve questions (four per skill) was presented to students in one of two possible linear presentations: blocked or interleaved. Students randomly assigned to the blocked condition received questions grouped by skill (A_1, A_2, A_3, A_4, B_1, B_2, B_3, B_4, C_1, C_2, C_3, C_4), while those randomly assigned to the interleaved condition received the same questions in a mixed skill pattern (A_1, A_2, B_1, B_2, C_1, C_2, A_3, B_3, C_3, B_4, C_4, A_4). All students partook in a follow-up assignment containing three questions (A_5, B_5, C_5) as a delayed posttest. The posttest was presented with tutoring in the form of on-demand hint messages and students were allowed multiple attempts to achieve accuracy.

The original work presented an Analysis of Covariance (ANCOVA) on the average posttest performance of 146 students (*n* Blocked = 60, *n* Interleaved = 86) based on binary scoring. Results only trended toward significance across the full sample, but split file analyses revealed significant learning gains for low skill students who had received the interleaved assignment. In a parallel analysis, average hint usage and attempt counts at posttest were considered through a Multivariate Analysis of Covariance (MANCOVA), with results suggesting a significant multivariate effect driven by a reduction in posttest hint usage for students in the interleaved condition. These results inspired the present work. Binary scoring alone could not consistently allow for reliable group differentiation until controlling for student skill level. Additionally, robust value was added via consideration of posttest variables that define partial credit in the present work. How would results have differed if the authors of the original work had considered algorithmic partial credit scoring?

3. METHODOLOGY

To examine the potential for using partial credit as a metric to more efficiently differentiate between groups, the dataset was processed using a definition of partial credit scoring previously validated within ASSISTments. Past research on modeling student performance within ASSISTments has revealed that certain definitions of partial credit significantly outperform others when attempting to predict next problem performance [6]. The algorithm presented in Figure 1, originally defined in [7], has been proven as an effective definition in the context of modeling student performance [7]. This algorithm establishes a score categorization based on logged information regarding the student's performance: the number of attempts required to reach an accurate response (attempt), the number of hints requested (hint_count), and whether or not the student was provided the answer through the bottom out hint (bottom_hint). A version of this algorithm was recently implemented within the ASSISTments platform.

After passing the dataset through the algorithm presented in Figure 1, the resulting file contained categorical partial credit scores (0, 0.3, 0.6, 0.7, 0.8, 1.0) for each students' performance on each problem in the practice and posttest sessions. Students could still earn full credit in the traditional sense (i.e., answering correctly on the first attempt), but only lost full credit if they made more than five attempts or were provided the answer through the bottom out hint. An example of the processed data, with variables from the original file as well as the resulting penalizations and partial credit scores, is presented in Table 1. The processed dataset has been stripped of student identifiers and is available at [11] for reference.

When considering user groups, this dataset offered two clear opportunities for group differentiation: experimental condition and discretized student performance level. The latter metric defines students as either high performing or low performing

```
IF attempt = 1 AND correct = 1 AND hint_count = 0
    THEN 1
ELSIF attempt < 3 AND hint_count = 0
    THEN .8
ELSIF (attempt <= 3 AND hint_count=0)
OR (hint_count = 1 AND bottom_hint != 1)
    THEN .7
ELSIF (attempt < 5 AND bottom_hint != 1)
OR (hint_count > 1 AND bottom_hint != 1)
    THEN .3
ELSE 0
```

Figure 1: Partial credit algorithm originally defined in [7].

Table 1: Randomized controlled trial data with partial credit algorithm employed

Student	Condition	Problem	Binary	Hints	Bottom Out	Attempts	Penalization	Partial Credit Score
Student 1	Interleaved	A_1	0	1	0	2	0.3	0.7
Student 1	Interleaved	B_1	0	0	0	2	0.2	0.8
Student 1	Interleaved	C_1	1	0	0	1	0.0	1.0
Student 2	Blocked	A_1	0	3	1	3	1.0	0.0
Student 2	Blocked	A_2	0	0	0	3	0.3	0.7
Student 2	Blocked	A_3	0	1	0	4	0.7	0.3

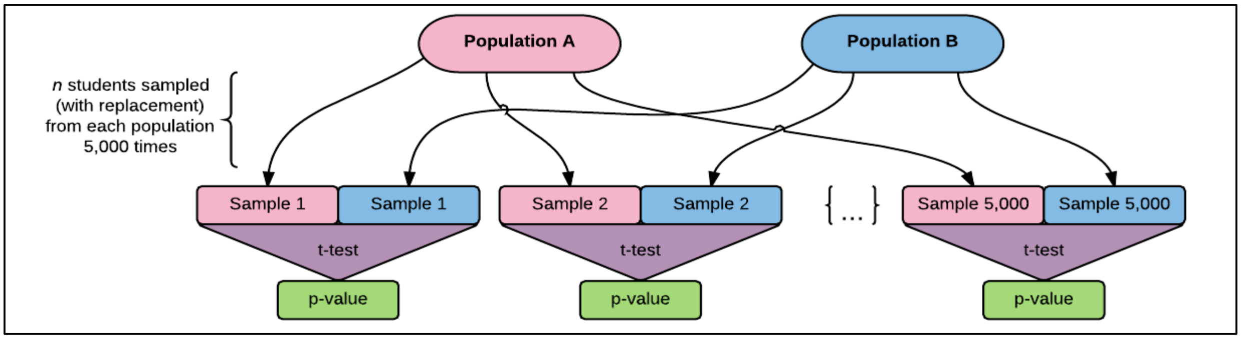

Figure 2: The resampling process used to create samples of *n* students from each population. Each set of samples was used in a t-test and significance values were recorded. This process was repeated 5,000 times for each group of *n* students.

based on a measure of prior knowledge calculated using the ASSISTments database. Prior knowledge is established by considering the average accuracy (in the binary sense) of all problems that a student has ever solved within ASSISTments. A median split can then be applied to this metric within a dataset to discretize groups of generally 'high performing' and generally 'low performing' students. In previous research, these groups have been found to exhibit significantly different performance, with low performing students logging reliably lower accuracy, more hints, and more attempts [8]. Thus, while observing differentiation between experimental conditions is subject to the success of the intervention, grouping students by skill level offers an obvious differentiation to test the efficacy of partial credit.

The full sample (146 students) was used to test differentiation between student performance levels. Equivalent samples of students were randomly selected from each performance level in single student increments (i.e., 5 students, 6 students, 7 students, etc.) For each set of equivalent samples of size *n*, an independent samples t-test was performed to compare the difference in partial credit scores between Sample 1 (a subset, *n*, of high performing students) and Sample 2 (a subset, *n*, of low performing students). A p-value denoting level of significance was recorded. This process was repeated to examine differences between Sample 1 and Sample 2 when considering binary scoring. These 'trials' were repeated 5,000 times per sampling increment. This process is depicted visually in Figure 2. For both partial and binary credit, sets of resulting p-values were then analyzed to determine the percentage of trials in which significant differences were observed between samples ($p < .05$). Findings were graphed for a visual comparison of the two scoring methods. Analyses and mappings were conducted using MATLAB [5] via code available for further consideration at [11].

This procedure was also used to differentiate between students based on experimental condition: blocked or interleaved. As the original work suggested that experimental condition only significantly altered achievement in low performing students, the present analysis considers only this subset of the original sample. Resampling with replacement was then used to establish artificial groups as large as desired. Please note that resampling is not employed in the present work to draw conclusions regarding the strength of a particular subsample or condition. The sole purpose of our analysis is to show that partial credit scoring can be used to reduce the sample sizes required to reliably differentiate between groups.

4. RESULTS

Results suggest that partial credit is exceptionally efficient in differentiating between distinct groups. Table 2 presents the differences in average correctness, hint usage, and attempt count observed when students are discretized into high and low performance levels - two groups that we know to be quite discernible and are therefore used here to validate our approach. Figure 3 depicts the percentage of samples in which significant differences were observed between these two groups. As these groups show obvious distinctions, both binary and partial credit scoring allow for 100% reliability of group differentiation with samples of fewer than 60 students. However, it should be noted that partial credit (red/dashed line) requires consistently smaller samples and attains reliability far more efficiently than binary scoring (blue/solid line). The resampling procedure suggested that

Table 2: Means and SDs for average correctness, hints, and attempts across performance levels

Group	Correctness	Hints	Attempts
Low Performing	0.54 (0.28)	0.72 (0.69)	2.05 (1.11)
High Performing	0.75 (0.22)	0.08 (0.21)	1.40 (0.43)

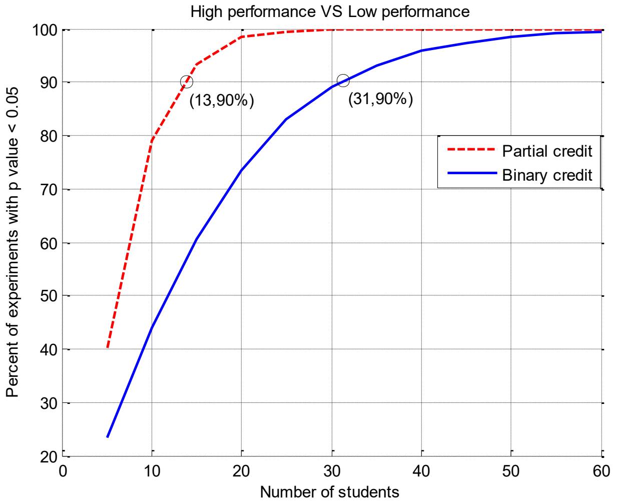

Figure 3: Significant differentiation in Performance Levels using Binary Scoring and Partial Credit Scoring. In groups with a known significant difference, differentiation is more efficient using partial credit. Sample size required for significant differentiation in 90% of trials is reduced by 58%.

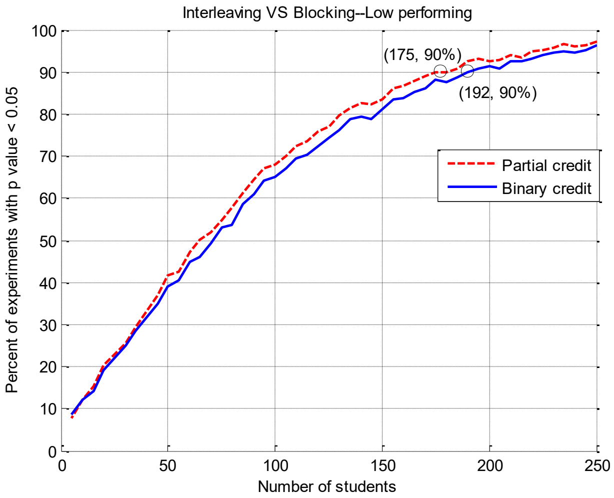

Figure 4: Significant differentiation in Condition using Binary Scoring and Partial Credit Scoring. In groups with a less substantial difference, differentiation is still almost always more efficient using partial credit. Sample size required for significant differentiation in 90% of trials is reduced by 9%.

Table 3: Means and SDs for average correctness, hints, and attempts across conditions for low performing students

Condition	Correctness	Hints	Attempts
Blocked	0.48 (0.25)	0.89 (0.67)	1.98 (0.58)
Interleaved	0.56 (0.29)	0.62 (0.67)	2.16 (1.37)

when using partial credit, equivalent groups of 13 students offer enough power to observe significant differences between performance levels in 90% of trials, while equivalent groups of 31 students were required when using binary scoring. Thus, within this context, using partial credit allowed sample sizes to be reduced by 58% while still obtaining the same result.

Although significant differences between experimental conditions within low performing students were more difficult to discern, as limited by the strength of the intervention, partial credit continued to offer more robust group differentiation when considering these user groups, as depicted in Figure 4. An analysis of means for the variables that combine to form partial credit revealed that low performing students in the interleaved condition were more accurate on average at posttest with fewer hints, as displayed in Table 3. Resampling suggested that when using partial credit, equivalent groups of 175 students offer enough power to observe significant differences between performance levels in 90% of trials, while equivalent groups of 192 students were required when using binary scoring. Thus, within this context, using partial credit allowed sample sizes to be reduced by 9% while obtaining the same result.

5. METHOD VALIDATION

When smaller equivalent sample sizes are required to differentiate between groups, Type II error is reduced for consistent sample sizes across scoring metrics. Before celebrating this finding, it is necessary to evaluate whether partial credit scoring in turn increases Type I error.

If no actual difference exists between two groups and we maintain a threshold of $p < .05$ in determining a significant difference, the Type I error rate, or alpha, should be 5%. In order to determine whether partial credit has reduced Type II error simply by increasing Type I error, we simulated a null experiment with our dataset. The full sample population (146 students) was subjected to the resampling (with replacement) process, without predefining students as having high or low performance or as belonging to a particular experimental condition. Thus, for every sample increment, *n*, Sample 1 and Sample 2 were randomly selected from the full population (establishing samples that were not distinctly different). An independent samples t-test was conducted to analyze the difference in partial credit scores between subsamples. This 'trial' was repeated 5,000 times, with p-values recorded for each trial. Complimentary trials were conducted using binary correctness. The percentage of trials resulting in significantly different subsamples is charted in Figure 5. Both measures show roughly similar trends, with approximately 5% of trials resulting in significant findings. This finding suggests that while partial credit allows for more robust group differentiation, it does not significantly influence Type I error.

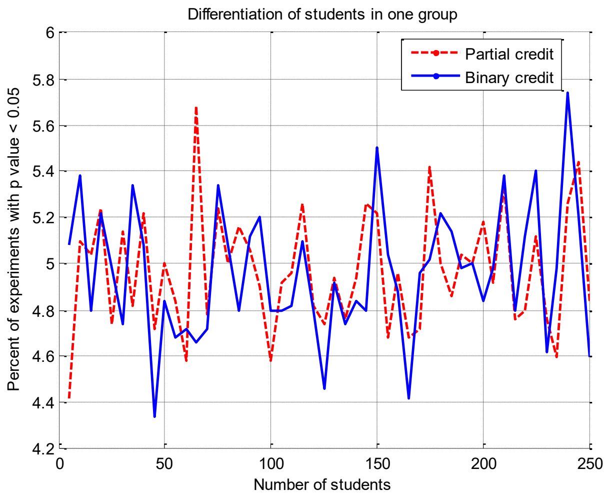

Figure 5: Type I error when resampling students from a solitary population using Binary Scoring and Partial Credit Scoring. Measures show roughly similar trends, suggesting that while partial credit allows for more robust group differentiation, it does not significantly impact Type I error.

6. DISCUSSION & FUTURE WORK

The present work sought to examine whether partial credit scoring could be used to enhance the efficiency of group differentiation within a previously published randomized controlled trial. Results confirmed our expectations, suggesting that partial credit is a more robust measure of student performance that increases the reliability of group differentiation and reduces the sample size required to observe significant differences (or, enhances power).

Partial credit scoring held merit for differentiating both between student performance levels and between experimental conditions. The lack of strength in the latter finding may be correlated with the efficacy of the intervention itself; differentiation based on a learning intervention should not be expected to be as robust as differentiation based on a mathematically established dichotomy. Still, trends in reliability for both scoring metrics follow the standards of a power analysis: if sample sizes in the original work had been larger, the intervention would have proven reliably significant.

It should be noted that while we observed consistent positive effects for partial credit, it *is* mathematically possible for the metric to underperform binary scoring. When using t-test comparisons, smaller p-values are obtained as t-statistics increase. T-statistics are inflated when mean differences between groups are large while variance within groups is low. Mathematically, the use of partial credit reduces within group variance while increasing the mean for each group. With this increase in means, it would be possible for binary scoring to outperform partial credit in a heavily skewed dataset.

A potential limitation of this approach can be found in the balance between enhancing group differentiation by adding measures of student performance and overfitting student performance. One could argue that to most efficiently differentiate between groups, all available student data could be collapsed into a partial credit metric, perhaps using a regression model. While this would likely result in better differentiation, the overly robust definition of 'partial credit' would fail to generalize to other online learning platforms, or possibly even to other content or user populations within the ASSISTments platform. Future work should consider the pros and cons of supplementing partial credit scoring with additional measures of student performance.

Another potential limitation of this work is that students' habits within the ASSISTments tutor are normative to those of a binary system; the majority of students understand that they will lose all credit if they request tutoring feedback or make more than one attempt. Thus, any definition of partial credit that uses a data mining approach to work backwards toward group differentiation should be considered potentially skewed. As partial credit was recently implemented within ASSISTments, future work should consider how the real-time effects of partial credit scoring impact the power of randomized controlled trials.

Future research should also consider how our partial credit approach contends with latent group differentiation, in an attempt to outperform modeling techniques like Knowledge Tracing. Even if latent, when two groups are qualitatively different (i.e., learned vs. unlearned, denoting skill mastery within KT) our method may be feasible to observe patterns leading to more reliable group differentiation. Future work should examine this paradigm, and consider the generalizability of using partial credit scoring within the context of other platforms and domains.

7. CONTRIBUTION

The work presented herein is novel in that it sought to bridge the gap between educational research and data mining by applying post hoc mining methods to the results of a previously published randomized controlled trial. Results suggested a substantial benefit of considering partial credit scoring within online learning platforms: increased efficiency in group differentiation which translates to increased power and reduced Type II error. Our findings further confirm the notion that allowing students to learn during assessment is beneficial to students and researchers alike. Student performance metrics that are typically lost on traditional posttests can actually improve data analysis. Further, our results suggest that by using robust measures of student performance, the number of items or opportunities analyzed need not be large to result in significant group differentiation, offering evidence for short, minimally invasive assessments. These findings translate to real world implications: significant outcomes can be observed with smaller samples and with fewer overall data points, reducing the many of the costs and constraints of experimental research.

8. ACKNOWLEDGMENTS

We acknowledge funding from multiple NSF grants (ACI-1440753, DRL-1252297, DRL-1109483, DRL-1316736, DRL-1031398), the U.S. Dept. of Ed. (IES R305A120125, R305C100024, GAANN), ONR, and the Gates Foundation.

9. REFERENCES

[1] Attali, Y. & Powers, D. (2010). Immediate feedback and opportunity to revise answers to open-end questions. *Educational and Psychological Measures*, 70 (1), 22-35.

[2] Corbett, A.T., Anderson, J.R. (1995). Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, 4, 253-278.

[3] Heffernan, N. & Heffernan, C. (2014). The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching. *Int J of AIED, 24(4),* 470-497.

[4] Koedinger, K.R. & Corbett, A.T. (2006). Cognitive tutors: Technology bringing learning science to the classroom. In K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (61-78). New York: Cambridge University Press.

[5] MATLAB version R.2013.a (2013). Natick, Massachusetts: MathWorks, Inc. Accessible at www.mathworks.com

[6] Ostrow, K., Donnelly, C., & Heffernan, N. (2015). Optimizing Partial Credit Algorithms to Predict Student Performance. In Santos, et al. (eds.) Proc of the 8th Int Conf on EDM, 404-407.

[7] Ostrow, K., Donnelly, C., Adjei, S., & Heffernan, N. (2015). Improving Student Modeling Through Partial Credit and Problem Difficulty. In Russell, D.M., Woolf, B., & Kiczales, G. (eds.) Proc of the 2nd ACM Conf on L@S, 11-20.

[8] Ostrow, K., Heffernan, N., Heffernan, C., Peterson, Z. (2015). Blocking vs. Interleaving: Examining Single-Session Effects within Middle School Math Homework. In Conati, Heffernan, Mitrovic, & Verdejo (eds.) Proc of the 17th Int Conf on AIED. Springer International, 388-347.

[9] Pardos, Z.A. & Heffernan, N.T. (2010). Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing. In De Bra, Kobsa, & Chin (eds.) Proc of the 18th Int Conf on UMAP, 255-266.

[10] Pardos, Z.A., & Heffernan, N.T. (2011). KT-IDEM: Introducing Item Difficulty to the Knowledge Tracing Model. In Joseph A. Konstan et al. (Eds.), Proc of the 19th Int Conf on UMAP, 243-254.

[11] Wang, Y. (2015). Data and Code for *Enhancing the Efficiency and Reliability of Group Differentiation through Partial Credit*: http://tiny.cc/LAK2016-Resampling

[12] Wang, Y. & Heffernan, N.T. (2011). The "Assistance" Model: Leveraging How Many Hints and Attempts a Student Needs. In Proc of the 24th Int FLAIRS Conf.

[13] Wang, Y. & Heffernan, N. (2013). Extending Knowledge Tracing to Allow Partial Credit: Using Continuous versus Binary Nodes. In K. Yacef et al. (Eds.) AIED 2013, LNAI 7926, 181-188.