(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

How Flexible Is Your Data? A Comparative Analysis of Scoring Methodologies across Learning Platforms in the Context of Group Differentiation

Korinn S. Ostrow
Worcester Polytechnic Institute, Worcester, MA, USA
ksostrow@wpi.edu

Yan Wang
Worcester Polytechnic Institute, Worcester, MA, USA

Neil T. Heffernan
Worcester Polytechnic Institute, Worcester, MA, USA

ABSTRACT. Data is flexible in that it is molded by not only the features and variables available to a researcher for analysis and interpretation, but also by how those features and variables are recorded and processed prior to evaluation. "Big Data" from online learning platforms and intelligent tutoring systems is no different. The work presented herein questions the quality and flexibility of data from two popular learning platforms, comparing binary measures of problem-level accuracy, the scoring method typically used to inform learner analytics, with partial credit scoring, a more robust, real-world methodology. This work extends previous research by examining how the manipulation of scoring methodology has the potential to alter outcomes when testing hypotheses, or specifically, when looking for significant differences between groups of students. Datasets from ASSISTments and Cognitive Tutor are used to assess the implications of data availability and manipulation within twelve mathematics skills. A resampling approach is used to determine the size of equivalent samples of high- and low-performing students required to reliably differentiate performance when considering each scoring methodology. Results suggest that in eleven out of twelve observed skills, partial credit offers more efficient group differentiation, increasing analytic power and reducing Type II error. Alternative applications of this approach and implications for the Learning Analytics community are discussed.

Keywords: Data flexibility, partial credit, group differentiation, resampling, ASSISTments, Cognitive Tutor

1 INTRODUCTION

1.1 The Overlooked Flexibility of Data

When analyzing a dataset, it is easy to fall back on the perspective that data collected in its natural or normalized form is all that is available for consideration. Researchers may not realize that their datasets are flexible; data can be manipulated or combined to consider different perspectives and, often, to

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

arrive at different conclusions. In some ways, arguments for data flexibility toe the line of how researchers can "lie with statistics," by using and abusing data cleaning, analysis, and presentation to sway an audience's interpretation (Huff, 1954). However, in the age of "Big Data," where data carries a high volume, variety, and velocity (Elgendy & Elragal, 2014), changing one's perspective of a dataset can be critically important. Big Data analysts define the processes by which data collection, storage, preparation, and analysis should be conducted, but often fail to consider how each of these steps interact to impact data quality (Janssen, van der Voort, & Wahyudi, 2017). Resulting data quality then influences statistical outcomes that drive decisions made from the data.

In the context of education, online learning platforms and intelligent tutoring systems have revolutionized the field's understanding of educational data by drastically expanding the amount of data available for mining and analytics. These systems allow for the collection of fine-grained data that can be aggregated in a variety of ways to examine student performance. However, no two systems are created equal; learning platforms can vary by domain, environment, and functionality, adding to the overall complexity and dynamic understanding of learner analytics. These differences make it difficult for researchers to establish broad, generalizable claims regarding best practices in technology-driven education.

Educational researchers should consider the flexibility of their data by assessing how sensitive available data is to cleaning, manipulation, and analytic efforts. The present work argues that educational data is flexible, molded by not only the variables available to a researcher for analysis and interpretation, but also by how those variables are processed prior to analysis.

1.2 Why Does Data Flexibility Matter?

How should researchers approach the problem of data flexibility? Is the problem a matter of data quality, potentially solved by refining data collection methodologies? Past work has shown that data accessibility and volume provide stronger links to resulting outcomes and decisions than data quality (O'Reilly, 1982; Keller & Staelin, 1987). Researchers may not even consider the *quality* of their data beyond traditional assumptions of normality or distribution, settling for that which is accessible. However, data quality should not be overlooked. Quality should be considered alongside the context of data collection and the researcher's personal decisions regarding analysis (O'Reilly, 1982).

When considered as a facet of data quality, perhaps the problem of data flexibility is actually a matter of data cleaning, approached by practices that alter the underlying distribution of critical variables (Dasu & Loh, 2012). While data cleaning is an important step in the process of analysis, over-cleaning may not establish more useful data, but instead may move data farther away from its real-world meaning (Dasu & Loh, 2012). Understanding the appropriate actions to take on a dataset is a skill linked to the data's context and accuracy (Morgan, 2015). The problem of data flexibility may also extend to feature generation and the legitimacy of using available data to derive new perspectives.

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

Although its definition may be varied, perhaps the most critical reason to consider data flexibility is its potential impact on statistical findings. Predictions and statistical outcomes are only as good as the input supplied for analysis. Decisions made from data extend from the processes by which the data was collected, processed, and analyzed (Janssen, van der Voort, & Wahyudi, 2017). In the context of educational research, the manner in which data is collected from online learning platforms and intelligent tutoring systems, and the ways in which that data is processed and interpreted ultimately have a profound impact on our views of effective educational practices.

1.3 Past Research and Present Goals

While a focus on data *quality* is hardly new, the idea of data *flexibility* is not often discussed. The present work considers a key question for Big Data analytics in education: what does data flexibility (i.e., the manipulation of available data and/or the use of feature generation to establish new data) mean for the evidence put forth in support of best practices in education? Specifically, this work examines how the nature of scoring employed by online learning platforms can alter results observed in the context of hypothesis testing.

In online learning platforms, students are typically asked to solve problems pertaining to particular skills or domains and receive feedback with regard to the accuracy of their responses. These platforms also commonly provide assistance in the form of hints, worked examples, scaffolding, or messages tied to specific wrong answers in order to help guide and improve student learning. Many online learning platforms rely on a binary scoring system that allots points based on the accuracy of the student's first response to each problem (i.e., a correct answer earns a score of 1, while an incorrect answer earns a score of 0).

This black and white depiction of students' problem-level understanding has progressed in recent years through research in the field of educational data mining. Mathematically speaking, a grey area, defined by a partial credit score, should usually paint a more robust picture of a student's knowledge. Intuition in the field of assessment has expanded to the world of online learning platforms, with evidence repeatedly observed for the importance of partial credit in technology-driven assessment contexts (Attali & Powers, 2008; 2010). Partial credit scoring offers a finer-grained distinction of student knowledge and increases the reliability of knowledge estimates (Attali, 2011). When data is subject to multiple interpretations (i.e., scoring patterns), it is critical to determine which format is more practical and credible in the real world (O'Reilly, 1982). Partial credit is not a novel concept, with teachers using various iterations of the practice in classrooms around the world. As such, the obvious next step is to establish credible partial credit metrics for online learning platforms, and subsequently, to gain a better understanding of how scoring methodology relates to student learning.

The present work seeks to provide empirical evidence for the use of partial credit scoring in online learning platforms based on a hypothesis testing approach. This work extends previous research on the use of partial credit scoring within the ASSISTments learning platform (Ostrow, Heffernan, Heffernan, &

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

Peterson, 2015; Ostrow, Donnelly, & Heffernan, 2015; Ostrow, Donnelly, Adjei, & Heffernan, 2015; Wang, Ostrow, Beck, & Heffernan, 2016) by examining how a system-specific definition of partial credit generalizes to data from a different platform.

One source of inspiration for this line of research is rooted in the field of educational data mining. Bayesian knowledge tracing (BKT), one of the most popular student modelling techniques, and that employed by Carnegie Learning's Cognitive Tutor, relies on the binary correctness of skill items (i.e., problems or questions pertaining to a particular skill) and a series of latent constructs to predict when a student will learn or "master" a skill (Corbett & Anderson, 1995). In recent years, researchers have tried to enhance BKT predictions by considering additional data pertaining to student performance. Essentially, by supplementing "black and white" performance data with additional features, researchers have established roundabout partial credit measures for gauging student performance. Features that have been considered in past work include personalized measures of students' prior knowledge (Pardos & Heffernan, 2010; Yudelson, Koedinger, & Gordon, 2013), estimates of item difficulty (Pardos & Heffernan, 2011), and algorithmically defined partial credit scores meant to replace binary scores (Wang & Heffernan, 2011; Wang & Heffernan, 2013; Ostrow, Donnelly, Adjei, & Heffernan, 2015). Despite myriad attempts to enrich student models through additional student and content data, it remains common practice for tutoring systems to report students' problem-level performance using binary accuracy.

A second source of inspiration for the present work is rooted in the field of educational psychology, specifically, in the analysis of randomized controlled trials. A randomized controlled trial examining the efficacy of interleaving skill content was conducted in ASSISTments. Published results revealed that significant differences between experimental and control conditions could only be observed when considering student performance data in addition to the binary accuracy of skill items (Ostrow, Heffernan, Heffernan, & Peterson, 2015). The partial credit of skill items, defined as a combination of the student's binary accuracy on an item and the number of hints and attempts used while solving the item, became a compelling research focus. Additional investigations have since revealed the benefits of partial credit when predicting student performance on future skill items (Ostrow, Donnelly, & Heffernan, 2015), pinpointed optimal system-specific definitions of partial credit (Ostrow, Donnelly, Adjei, & Heffernan, 2015), and shown that partial credit scoring can lead to a more robust understanding of student knowledge (Desmarais & Baker, 2011).

To test an algorithmic definition of partial credit designed specifically for the ASSISTments platform, data from Ostrow, Heffernan, Heffernan, and Peterson's (2015) randomized controlled trial was revived and reanalyzed. Wang et al. (2016) confirmed that when comparing the dataset's original binary item scoring to partial credit scores derived from an algorithm that combined penalties for hint and attempt usage, significant differences between experimental and control conditions could be found more efficiently. Enhanced efficiency was defined as requiring smaller sample populations to reliably observe group differences within hypothesis tests, essentially equating to an increase in power and a reduction

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

of Type II error (Wang et al., 2016). The work also examined the use of partial credit when differentiating between groups of students *known* to differ significantly: high-performing and low-performing students. By comparing groups with a *known* difference to groups from an experimental intervention with an unknown effect size, the authors observed that the efficiency gained using partial credit was potentially mediated by the magnitude of the true difference between groups (Wang et al., 2016). That is, groups *known* to differ significantly, established using students' knowledge estimates measured prior to randomization, saw increased efficiency in group differentiation more dramatic than that observed for groups differing by an unknown magnitude.

The present work seeks to extend past research by examining the paradigm of enhanced group differentiation through partial credit scoring in non-experimental contexts using mastery-based learning assignments within ASSISTments. Additionally, the generalization of this approach is considered by examining similar skill content sourced from Cognitive Tutor. It was hypothesized that, as previously observed in Wang et al. (2016), partial credit scoring would enhance the efficiency with which significantly different groups of students could be detected. Specifically, this work examines groups of students *known* to differ significantly: high-performing and low-performing students. Support for this hypothesis would emphasize the notion that researchers should consider the flexibility of their data, especially in the context of hypothesis testing and between-group analyses.

2 DATASETS

The flexibility of data from two popular learning platforms, ASSISTments[1] and Carnegie Learning's Cognitive Tutor,[2] is considered in the present work. These systems are used to examine how the nature in which students are scored on skill items (i.e., problems or questions pertaining to a particular skill) can sway the results of analyses.

2.1 ASSISTments

Past research on partial credit and group differentiation conducted in an experimental setting within ASSISTments served as inspiration for the present work. The ASSISTments platform has a primary focus in K–12 mathematics, supporting over 50,000 students around the world as they complete classwork and homework. Delivering textbook content and skill-based assignments aligned with the Common Core State Standards (NGACBP & CCSSO, 2010), students gain insight from immediate feedback and a variety of rich tutoring strategies, while teachers receive powerful student- and class-level reports that support opportunities for formative assessment (Heffernan & Heffernan, 2014). The ASSISTments platform also strives to advance educational research, serving as a tool for researchers to embed randomized controlled trials meant to examine the effects of learning interventions at scale (Ostrow & Heffernan, 2016).

[1] http://www.assistments.org/

[2] http://www.carnegielearning.com/

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

The ASSISTments dataset used in the present work is composed of "Skill Builder" data accrued between September 2009 and December 2014. Skill Builders are mastery-based learning assignments mapped to the Common Core State Standards (NGACBP & CCSSO, 2010), and are commonly assigned by teachers using ASSISTments. These assignments provide a wealth of data for learning analytics research. Within a Skill Builder, students must complete a series of problems randomly selected from a skill pool until meeting a predefined threshold of skill "mastery" (i.e., answering three consecutive problems, or skill items, accurately on the first attempt). While working through a Skill Builder, students can request hints and make multiple solution attempts, with all performance data logged by the system. For brevity, the dataset considered herein examines the six most commonly assigned Skill Builders within ASSISTments from 2009–2014 (details presented in Table 1). This is non-experimental data; the Skill Builders considered were not altered to include experimental interventions during the collection period. Although other skill topics could have been considered, these six topics offered the highest volume of student data per skill, and as the volume of data degrades, so too would the quality of analysis. Therefore, these six skills are presented as a proof of concept for the analysis presented herein.

Table 1: Details Pertaining to Analyzed ASSISTments Skill Builders

Skill Topic	Students	Difficulty
Equation Solving, Two or More Steps	5,269	0.43
Greatest Common Factor	5,169	0.42
Distributive Property	5,693	0.37
Multiplying Fractions and Mixed Numbers	4,719	0.26
Positive and Negative Integers	6,314	0.20
Scientific Notation	6,502	0.19

Note. Difficulty increases toward 1.0.

For each Skill Builder, an estimate of difficulty was calculated by averaging accuracy across all items within the skill. The inverse of this measure was then taken to establish a scale in which difficulty increases toward 1.0. Thus, a difficulty score of 0.37 would mean that students averaged 63% accuracy across all problems in the skill, while a difficulty score of 0.19 would mean that students averaged 81% accuracy across all problems in the skill. Level of difficulty was considered in an attempt to assess whether difficulty hinders the power of partial credit for group differentiation (perhaps by altering the number of hints or attempts required by students and thereby suggesting sensitivity in how partial credit metrics should be defined). Within Table 1, the Skill Builders are presented from most difficult (equation solving with more than two steps) to least difficult (scientific notation).

Additionally, all skill items within the ASSISTments dataset carried binary scores for student performance, based on the student's first action or solution attempt. The dataset also quantified the number of hints and attempts each student required per skill item. This extra data is algorithmically combined with student accuracy to establish partial credit scores, as discussed in the Methods section.

The analyses presented herein consider a student's average performance across the first three items solved in each Skill Builder. For instance, when considering two students solving problems pertaining to

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

the distributive property, Susie displayed the binary solution pattern {inaccurate, inaccurate, accurate, or 0, 0, 1} on the first three skill items, while John displayed the binary pattern {accurate, inaccurate, accurate, or 1, 0, 1}. By isolating the first three skill items, or problems solved, this work attempts to prove that group differentiation can be accomplished more efficiently through partial credit even when student performance data is strictly limited. Using this approach, the dataset was cleaned to remove students answering fewer than three items within each skill.

2.2 Cognitive Tutor

A Cognitive Tutor dataset was considered to assess how the algorithm for partial credit, originally defined within the context of ASSISTments, generalized to a different platform. Cognitive Tutor bears some similarity to ASSISTments, but it also differs in several ways. Cognitive Tutor is a series of broad-reaching tutoring systems for students in grades 9–12 distributed by Carnegie Learning (2016). The Cognitive Tutor series is built around the ACT-R theory of cognition, enlisting problem-solving techniques to compare automated solution steps against student solutions and provide immediate feedback and assistance as necessary (Anderson, Corbett, Koedinger, & Pelletier, 1995; Ritter, Anderson, Koedinger, & Corbett, 2007). Cognitive Tutor is distributed as a portion of broader curriculum reform, with courses available in multiple mathematics domains (VanLehn et al., 2005; Carnegie Learning, 2016). Schools or districts adopt Cognitive Tutor and teachers assign tutored content for classwork or homework in alignment with other Carnegie Learning materials.

The Cognitive Tutor dataset used in the present work is comprised of data from the Algebra 1 course that was promoted as part of the Knowledge Discovery and Data Mining (KDD) Cup dataset in 2010 (KDD Cup, 2010). The full dataset, as retrieved from the PSLC DataShop (Stamper et al., 2011), spans a single academic year (2005–2006), with over 880K skill items completed by 559 students working within 106 Algebra skills. Again, for brevity and as a proof of concept, the present investigation focuses on data from the six skills exhibiting the highest volume of student data. Details pertaining to these six skills are presented in Table 2.

Table 2: Details Pertaining to Analyzed Cognitive Tutor Skills

Skill Topic	Students	Difficulty
Expressions with Negative Slopes	263	0.68
Combining Like Terms	264	0.38
Find X with Positive Slopes	268	0.35
Labelling Axes	263	0.33
Consolidate Variables with Coefficients	266	0.15
Consolidate Variables without Coefficients	263	0.10

Note. Difficulty increases toward 1.0.

Estimates of difficulty for each of the six skills were again calculated by averaging students' accuracy across all items within the skill and subtracting that number from 1. Within Table 2, skills are presented from most difficult (Expressions with Negative Slopes) to least difficult (Consolidate Variables without Coefficients).

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

All skill items within the Cognitive Tutor dataset carried binary scores for student performance, based on the student's first action or attempt (paralleling the format of the ASSISTments dataset). The dataset also quantified the number of hints and attempts required by each student within each skill item. Paralleling the preprocessing method used to clean the ASSISTments dataset, the Cognitive Tutor dataset was cleaned to remove students answering fewer than three items within each skill. Isolating the first three skill items within each skill then offered a comparable view of the efficacy of the approach given limited student data in an alternative platform.

3 METHODS

As the present work seeks to extend previous research on the efficiency and reliability of partial credit in the context of group differentiation, the methodology presented herein was adapted from previous work and uses much of the same terminology (Wang et al., 2016). The following subsections highlight the three primary steps required to evaluate partial credit in the context of group differentiation.

3.1 Appending Partial Credit Scores

The present analysis relies on an algorithmic definition of partial credit adapted[3] from that originally presented in Wang et al. (2016) and Ostrow, Donnelly, Adjei, and Heffernan (2015) (see Figure 1). For each skill item or problem solved, the algorithm considered the student's binary score (*correct*), the number of attempts required to solve the item (*attempt*), the number of hints required to solve the item (*hint_count*), and a binary flag denoting whether the student was shown the answer through a "bottom out hint" (*bottom_hint*).

```
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 (attempt <=3 AND hint_count = 1 AND bottom_hint != 1)
   THEN .7

ELSIF (attempt < 5 AND bottom_hint != 1)
  OR (attempt < 5 AND hint_count > 1 AND bottom_hint != 1)
   THEN .3

ELSE 0
```

Figure 1: Partial credit algorithm adapted from Ostrow, Donnelly, Adjei, and Heffernan, 2015.

[3] It should be noted that this definition of partial credit was designed for data mining within ASSISTments. As such, two points in the original code were found to be susceptible to outlying student behaviours in a new context. In the present work, the definitions for arriving at scores of 0.7 and 0.3 have been amended to remove this potential flaw and to extend the capacity for generalization to other platforms. These flaws would have allowed students to make an endless number of attempts as long as the hint count requirement was met for scores of 0.7 and 0.3. Datasets for previous work using the original pseudo code were reviewed and no such outliers existed, leaving published results unaltered. The errors only became apparent when applying the code to data from a new platform, thus requiring amendment for the present work.

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

The algorithm presented in Figure 1 was applied to both datasets and categorical partial credit scores (0, 0.3, 0.7, 0.8, 1.0) were appended to each skill item for each student. Domain experts and the creators of ASSISTments established these values following a data mining endeavor that examined the predictive ability of various penalizations for hint and attempt usage in the system (Ostrow, Donnelly, & Heffernan, 2015). Examples of resulting partial credit scores appended to ASSISTments and Cognitive Tutor data are presented in Table 3. Full versions of the modified datasets have been stripped of student identifiers and are available in Wang (2016).

Table 3: Excerpts from ASSISTments and Cognitive Tutor Datasets Exemplifying Partial Credit Scoring

Student/ System	Performance	Skill	Opportunity	Binary	Hints	Attempts	Bottom	Partial Credit Score
1-ASM	High	Distributive Property	1	0	1	2	0	0.7
1-ASM	High	Distributive Property	2	0	0	2	0	0.8
1-ASM	High	Distributive Property	3	1	0	1	0	1.0
2-ASM	Low	Scientific Notation	1	0	2	3	0	0.3
2-ASM	Low	Scientific Notation	2	0	3	3	1	0.0
2-ASM	Low	Scientific Notation	3	1	0	1	0	1.0
1-COG	Low	Combine Like Terms	1	0	3	4	1	0.0
1-COG	Low	Combine Like Terms	2	0	0	3	0	0.7
1-COG	Low	Combine Like Terms	3	0	3	2	0	0.3
2-COG	High	Labelling Axes	1	1	0	1	0	1.0
2-COG	High	Labelling Axes	2	1	0	1	0	1.0
2-COG	High	Labelling Axes	3	0	1	2	0	0.7

Note. ASM = ASSISTments, COG = Cognitive Tutor. Performance = Discretized student performance level established from data prior to student participation in the considered skill. Opportunity = Sequential count of skill items experienced. Binary = Original binary score. Hints, Attempts, and Bottom flag = student performance metrics for use in calculating partial credit.

3.2 Discretizing Student Performance

Within each dataset, all student accuracy data collected *prior* to the student's involvement with the six Skill Builders or the six Algebra 1 skills used in the present analyses was averaged to estimate a level of global knowledge or performance. Specifically, this approach aggregates the accuracy data for all items across all skills with earlier time stamps than those reflecting student participation in the skills considered in the present analysis. In order to establish groups with a *known* difference for the present analysis, dichotomization of the resulting variable around the median can then be used to form two groups of distinctly different students: high-performing students (those typically answering problems with high accuracy) and low-performing students (those typically answering problems with low accuracy). Table 3 includes a target flag denoting student performance level, derived from this process. The amount of historical student performance data was more prevalent for the ASSISTments dataset given its span of five academic years. Historical student-performance data in the Cognitive Tutor dataset was limited by data availability, considering its span of only a single academic year, and as such, it is possible that student performance classifications are more accurate in the context of ASSISTments.

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

Discretized student performance was used in the present work to isolate groups of students *known* to have significantly different performance. Low-performing students tend to exhibit reliably lower accuracy, higher hint usage, and higher attempt counts (Ostrow, Heffernan, Heffernan, & Peterson, 2015). This *known* dichotomy offers a ground truth from which to examine the strength of partial credit scoring for group differentiation. If the difference between groups is known, the question becomes "How efficiently can the difference be detected?" Validating this method on mathematically distinct groups supports the later use of partial credit in the analysis of randomized controlled trials in which the magnitude of group difference is unknown and subject to the effect size of a learning intervention.

3.3 Resampling with Replacement

After appending partial credit scores and discretizing students by performance level, the datasets were primed for examining the efficiency of partial credit in comparison to binary scoring through a rigorous resampling procedure (depicted visually in Figure 2). Equal sized groups of students were randomly sampled (with replacement) from discretized performance levels in increments of five students (i.e., 5 students, 10 students, 15 students, etc.). The replacement procedure allowed equivalent sample sizes to extend beyond the actual number of students available in the dataset to examine the simulated efficacy of partial credit within larger samples following the same distributions.

Following each paired sampling, an independent samples t-test was conducted to compare the difference in *partial credit scores* between *performance levels*. A second independent samples t-test was conducted to compare the difference in *binary credit scores* between *performance levels*. Both resulting p-values were recorded, thereby concluding a single "trial." These "trials" were repeated 5,000 times per sampling increment, *n*. This process produced a list of 5,000 p-values per scoring style, per equivalent sampling increment. These lists were then assessed to determine the percentage of trials in which differences between student performance levels were observed to be significant ($p < .05$) for each sampling increment, *n*. Findings for each scoring style were graphed for comparison, and are presented in Section 4, Figures 3 and 4. All analyses and mappings were conducted using MATLAB (2013) via code available in Wang (2016).

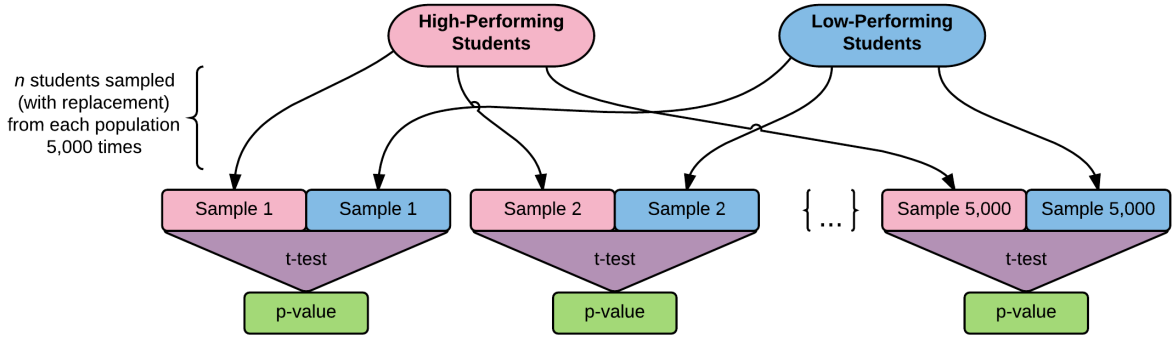

Figure 2: The resampling (with replacement) process from Wang et al. (2016).

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

4 RESULTS

4.1 ASSISTments

Considering the ASSISTments dataset, results suggested that partial credit consistently offered more efficient group differentiation. For each Skill Builder, an analysis of means was conducted to compare average binary correctness, hint usage, and attempt count within the first three skill items (or problems) experienced by each student. Results depicted distinct trends between discretized performance levels, as shown in Table 4. High-performing students typically showed higher average correctness, with a lower prevalence of hint usage and a lower average attempt count than low-performing students.

Table 4: Means and SDs for Correctness, Hints, and Attempts across Student Performance Levels in Analyzed ASSISTments Skill Builders

Skill Topic	Correctness	Hints	Attempts
Equation Solving, Two or More Steps			
High	0.65 (0.33)	0.63 (0.83)	1.82 (3.80)
Low	0.49 (0.37)	1.13 (1.05)	2.04 (2.46)
Greatest Common Factor			
High	0.65 (0.30)	0.42 (0.68)	1.95 (6.24)
Low	0.50 (0.33)	0.94 (0.95)	2.56 (3.16)
Distributive Property			
High	0.71 (0.31)	0.47 (0.80)	1.77 (2.93)
Low	0.55 (0.35)	0.93 (1.04)	2.14 (4.08)
Multiplying Fractions and Mixed Numbers			
High	0.82 (0.25)	0.22 (0.50)	1.72 (10.22)
Low	0.66 (0.32)	0.67 (0.89)	1.91 (2.96)
Positive and Negative Integers			
High	0.87 (0.22)	0.08 (0.30)	1.24 (0.56)
Low	0.73 (0.31)	0.26 (0.62)	1.66 (2.27)
Scientific Notation			
High	0.86 (0.23)	0.13 (0.40)	1.33 (1.01)
Low	0.75 (0.30)	0.35 (0.71)	1.83 (6.36)

Table 5: Group Size at which 90% of Samples Result in Significant Differentiation ($p < .05$) for Analyzed ASSISTments Skill Builders

	Group Size		Efficiency Gained
Skill Topic	Partial	Binary	Binary to Partial
Equation Solving, Two or More Steps	75	95	21%
Greatest Common Factor	50	90	44%
Distributive Property	85	100	15%
Multiplying Fractions and Mixed Numbers	55	70	21%
Positive and Negative Integers	70	85	18%
Scientific Notation	115	140	18%
Note. A paired samples t-test of group sizes suggested that observed sample reductions were significant, $p < .01$.			

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

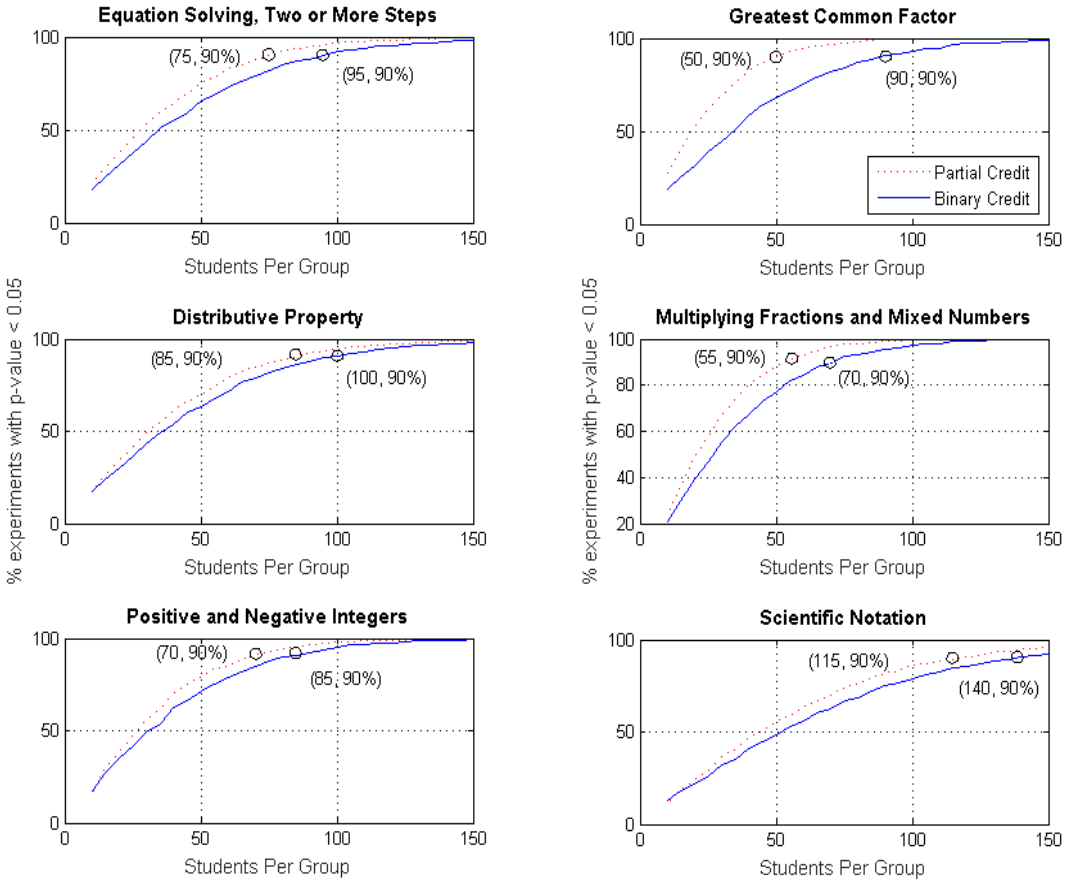

Figure 3: Significant differentiation of student performance level across six ASSISTments Skill Builders using binary scoring (solid blue) and partial credit scoring (dashed red). Sample size required for significant differentiation in 90% of trials was reduced by an average of 22.9% (SD = 10.8%, Range = 15–44%).

The set of graphs in Figure 3 reflect the percentage of trials in which significant differences ($p < .05$) were observed between performance levels for each Skill Builder and each style of scoring. Skill Builders are presented from most difficult (top left) to least difficult (bottom right). For all graphs, dashed red lines denote partial credit and solid blue lines denote binary scoring. Within each Skill Builder, partial credit consistently outperformed binary scoring across sampling increments. Partial credit allowed reliable group differentiation to be attained with significantly fewer students regardless of Skill Builder topic. A paired samples t-test of group sizes suggested that observed sample reductions were significant, $p < .01$. The magnitude of efficiency gained was determined by calculating and comparing the size of equivalent samples required to observe significantly different groups in 90% of trials. That is to say, the requirement was a significant difference ($p <. 05$) in at least 90% of trials for a given sample size, n. This magnitude differed across sets, but did not appear to be correlated with skill difficulty. Efficiency gained is pinpointed in each graph within Figure 3, and presented in detail in Table 5. The average efficiency gained by partial credit scoring across Skill Builders was 22.9%, with a standard deviation of 10.8%.

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

4.2 Cognitive Tutor

Considering the Cognitive Tutor dataset, results suggested that partial credit offered more efficient group differentiation in five out of the six skills analyzed. For each skill, an analysis of means was conducted to compare average binary correctness, hint usage, and attempt count within the first three items (or problems) experienced by each student. Results again depicted highly discretized performance levels, as shown in Table 6. High-performing students again showed higher average correctness, with a lower prevalence of hint usage and a lower average attempt count than low-performing students.

Table 6: Means and SDs for Correctness, Hints, and Attempts across Student Performance Levels in Analyzed Cognitive Tutor Skills

Skill Topic	Correctness	Hints	Attempts
Expressions with Negative Slopes			
High	0.42 (0.31)	1.10 (1.36)	2.41 (1.27)
Low	0.26 (0.30)	2.01 (1.85)	2.91 (1.72)
Combining Like Terms			
High	0.72 (0.30)	0.18 (0.53)	3.64 (3.26)
Low	0.53 (0.35)	0.46 (1.09)	5.05 (4.49)
Find X with Positive Slopes			
High	0.72 (0.27)	0.46 (1.10)	1.93 (2.21)
Low	0.58 (0.28)	1.35 (1.97)	2.57 (1.95)
Labelling Axes			
High	0.69 (0.30)	0.17 (0.49)	1.38 (0.51)
Low	0.65 (0.32)	0.38 (0.98)	1.45 (0.63)
Consolidate Variables with Coefficients			
High	0.88 (0.22)	0.08 (0.29)	1.18 (0.35)
Low	0.81 (0.25)	0.23 (0.55)	1.30 (0.55)
Consolidate Variables without Coefficients			
High	0.92 (0.20)	0.05 (0.20)	1.09 (0.34)
Low	0.88 (0.26)	0.10 (0.32)	1.12 (0.31)

The set of graphs in Figure 4 reflect the percentage of trials in which significant differences ($p < .05$) were observed between performance levels for each skill and each style of scoring. Again, skills are presented from most difficult (top left) to least difficult (bottom right), with dashed red lines denoting partial credit and solid blue lines denoting binary scoring. Partial credit failed to outperform binary credit in one skill, Combining Like Terms. Within this skill, binary credit outperformed partial credit, reaching reliable group differentiation with equivalent samples of 60 students, while partial credit required equivalent samples of 85 students (a 42% increase in sample size). This skill may have been fundamentally different, as in all other skills, partial credit allowed reliable group differentiation to be attained with significantly fewer students. Note that the number of students required to observe significant differences in 90% of trials differed across skills, thus causing variation in the x-axis across graphs, an unnecessary requirement in the context of ASSISTments Skill Builders. The magnitude of efficiency gained was differential across sets but did not appear to correlate with skill difficulty. Efficiency gained is pinpointed in each graph within Figure 4, and presented in detail in Table 7. Considering the five skills in which group differentiation was made more efficient through partial credit

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

scoring, the average efficiency gained across skills was 41.2%, with a standard deviation of 16.4%. When all six skills were examined together, the average efficiency gained across skills dropped to 27.4%, with a standard deviation of 36.9%. Interestingly, while efficiency gained across skills offered a significant improvement for group differentiation within ASSISTments data, the same was not true for Cognitive Tutor skills; a paired samples t-test of group sizes suggested that observed sample reductions were not significant, $p > .05$.

Table 7: Group Size at which 90% of Samples Result in Significant Differentiation ($p < .05$) for Analyzed Cognitive Tutor Skills

	Group Size		Efficiency Gained
Skill Topic	Partial	Binary	Binary to Partial
Expressions with Negative Slopes	60	85	29%
Combining Like Terms	85	60	–42%
Find X with Positive Slopes	55	75	27%
Labelling Axes	430	1055	59%
Consolidate Variables with Coefficients	115	280	59%
Consolidate Variables without Coefficients	385	565	32%

Note. A paired samples t-test of group sizes suggested that observed sample reductions were not significant, $p > .05$.

5 METHOD VALIDATION

To mirror the validation check in previous work (Wang et al., 2016), a final analysis was conducted to verify that the observed reduction in Type II error made possible by partial credit (i.e., smaller sample sizes required to differentiate between discretized groups) was not linked to an inflation in Type I error.

When groups are *not* significantly different (i.e., homogenous, random groups), Type I error should equal 5% (i.e., the alpha value). To verify this concept, null trials were simulated for each Skill Builder and Cognitive Tutor skill by randomly selecting students from respective datasets (disregarding performance level) to establish two homogenous groups of students with no anticipated group difference. P-values were collected from 5,000 trials for each scoring metric, following the resampling method presented in Section 3.3, and results were graphed. As all cases resulted in similar graphs, one exemplary skill was chosen from each system for presentation in Figure 5. Results showed similar and nondescript noise around the alpha value of .05, suggesting that while partial credit allowed for more efficient group differentiation, it did not significantly inflate Type I error, confirming findings from previous work (Wang et al., 2016).

As also noted in Wang et al. (2016), a known limitation of this work is that it is *mathematically possible* for partial credit to underperform binary scoring. This may be why binary credit was observed to be more efficient at differentiating between student performance levels in the context of the Cognitive Tutor skill "Combining Like Terms." When using t-test comparisons, smaller p-values are obtained as the t-statistic increases. T-statistics inflate as mean differences between groups increase while variance within groups decreases. Mathematically, the use of partial credit reduces within-group variance while inflating the group mean, simultaneously narrowing and shifting distributions to the right. Despite an increase in means, it is possible for binary scoring to outperform partial credit in a skewed dataset. For

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

instance, Table 8 examines two examples in which between-group (A & B) comparisons of scoring metrics are assessed using independent samples t-tests, similar to a single "trial" from the resampling approach used in the present work. Note that the scoring methodology in this example is not restricted to the partial credit values used in the present work, but is merely offered as a mathematical proof that partial credit is not *always* more effective. Although Example 1 looks quite similar to the majority of the findings presented herein, Example 2 reveals a scenario in which binary scoring outperforms partial credit, resulting in a lower p-value, similar to that observed in the "Combining Like Terms" Cognitive Tutor skill.

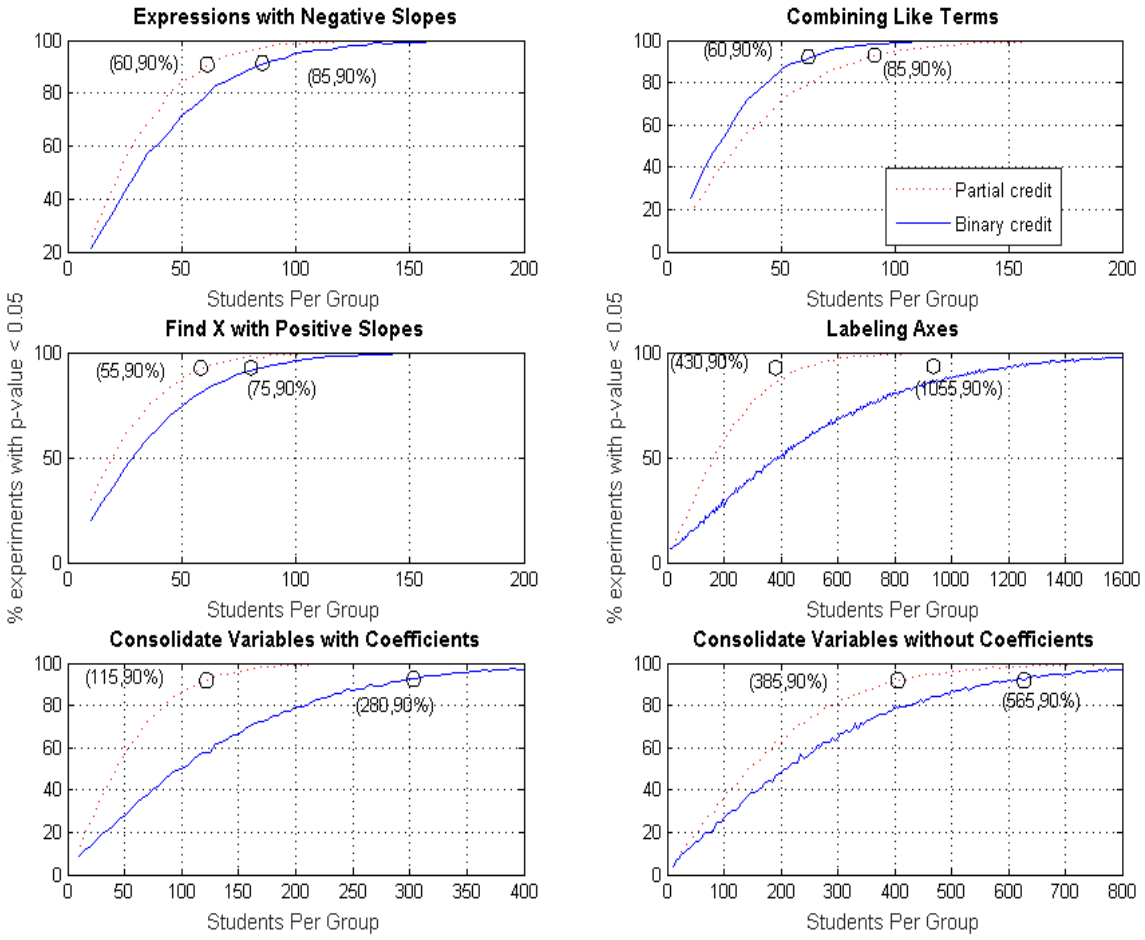

Figure 4: Significant differentiation of student performance level across six Cognitive Tutor skills using binary scoring (solid blue) and partial credit scoring (dashed red). Among most skills (5/6), sample size required for significant differentiation in 90% of trials was reduced by an average of 41.2% (SD = 16.4%, Range = 27–59%). Binary credit was found to be more successful at differentiating between groups in the skill "Combining Like Terms," swaying the effect of the data manipulation when considering all skills (M = 27.4%, SD = 36.9%).

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

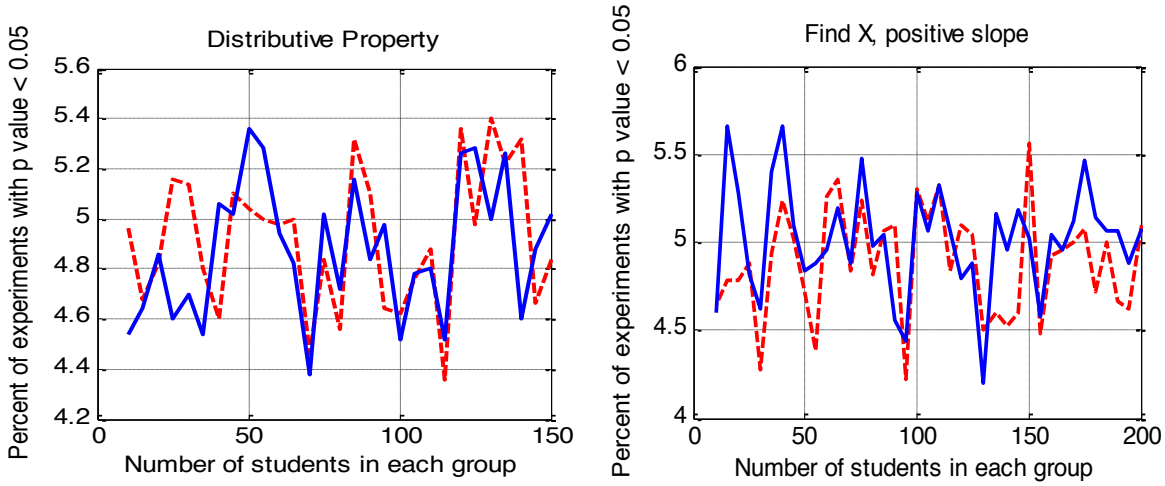

Figure 5: Type I error within an ASSISTments Skill Builder (left) and a Cognitive Tutor skill (right) using binary scoring (solid blue) and partial credit scoring (dashed red). These two measures show natural noise around the alpha value, α = 0.05, suggesting that while partial credit typically allowed for more robust group differentiation, it did not significantly influence Type I error.

Additionally, sample size and relative sample homogeneity may have impacted the results observed across skills. Given the rigorous resampling technique applied, particular skills (especially in the context of Cognitive Tutor) required a greater number of students to detect significant group differences in 90% of trials. This suggests high homogeneity of performance within the skill (i.e., little variance in the number of hints or attempts used by students regardless of performance level). Future work should consider how a partial credit definition can be tailored to specific skills to better discretize or model performance.

Table 8: Potential for Partial Credit to Outperform Binary Scoring (Example 1) and Reverse (Example 2)

Example 1				Example 2			
Binary Scoring		*Partial Credit*		*Binary Scoring*		*Partial Credit*	
A	B	A	B	A	B	A	B
1	1	1	1	1	1	1	1
1	0	1	0.1	1	0	1	0.6
1	1	1	1	1	1	1	1
0	0	0.5	0.3	0	0	0.2	0.3
0	0	0.8	0.2	0	0	0.5	0.2
0	0	0.3	0.8	0	0	0.3	0.8
$t = 0.53$		$t = 0.96$		$t = 0.53$		$t = 0.08$	
$p = 0.60$		$p = 0.36$		$p = 0.60$		$p = 0.94$	

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

6 DISCUSSION

This work extended previous research on the power of partial credit scoring in the context of group differentiation (Wang et al., 2016). Using mastery learning datasets from ASSISTments and Cognitive Tutor, algorithmically defined partial credit was compared to traditional binary scoring for detecting significant differences between discretized student performance levels. Results suggested that in eleven out of twelve skills analyzed, partial credit proved more efficient than binary scoring, requiring smaller samples to reach reliably significant group differentiation. This means that by considering a more robust perspective of student performance, it was possible to increase the power of analyses and reduce Type II error without increasing Type I error. This translates to data flexibility in that the observed outcomes at particular sample sizes, *n*, could differ based on the amount, type, and quality of data considered.

The results of the present work suggest that when exploring the differences between two sample populations, perhaps most importantly in the context of determining the strength of educational interventions, it is critical to consider all available data and ask questions of the data's quality. The analysis presented here has shown that changing the scoring methodology of online learning platforms by considering additional student data can alter the results observed at differing sample sizes, *n*. For instance, with a smaller sample size and reduced power, a researcher may conclude that a specific learning intervention did not significantly improve student learning (i.e., the experimental group did not differ significantly from the control group) based on an examination of binary problem accuracy. By considering more robust information about the students in the sample, the researcher may uncover value in the intervention that was otherwise hidden at the given sample size, as observed by Ostrow, Heffernan, Heffernan, and Peterson (2015).

As such, the present findings confirm the notion that allowing students to learn during assessment can be beneficial to students and researchers alike. As observed in related work, the reliability of knowledge estimates increases using partial credit within assessment environments (Attali, 2011). Rich features that are often lost on formal post-tests can help to better define student performance, allowing for improved data analysis. Further, these 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 may impact student modeling conducted by those in the learning analytics community.

For each platform, the observed efficiency gains in group differentiation varied by skill content but did not appear to be directly linked to skill difficulty. Instead, it is possible that differences observed in the amount of efficiency gained were linked to the algorithm used for partial credit calculation. As the algorithm was initially conceived by domain experts for data mining ASSISTments log files (Ostrow, Donnelly, Adjei, & Heffernan, 2015), it is important to note that while the potential for generalization to other systems with similar data was observed, tailoring the definition of partial credit to other systems, or establishing skill specific algorithms would likely produce better fitting models. Data mining can be used to observe how students interact with an online learning platform, and to gain insight regarding

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

how those interactions should influence system or skill specific partial credit scoring. For instance, previous work within ASSISTments has shown that partial credit scores were more sensitive to attempt penalizations than hint penalizations (Ostrow, Donnelly, & Heffernan, 2015), although attempt count may not help to significantly differentiate between groups within the system (Ostrow, Heffernan, Heffernan, & Peterson, 2015). While Cognitive Tutor logs many of the same student/system interactions, the emphasis of attempts and tutoring usage on optimal scoring could vary greatly in this different context. Future work should further examine the sensitivity of how partial credit is defined and how the generalizability of a static algorithm could be improved when considering similar systems.

Similarly, a potential limitation of this approach is the balance between enhancing group differentiation through rich features and the potential for overfitting student performance. One could argue that to most efficiently differentiate between groups, all available student data could be considered. However, while this would result in near perfect differentiation, the overly robust view of performance would fail to generalize to other intelligent tutoring systems, or possibly even to other content within the same platform. Results suggest that overfitting may have been a factor in the present analysis, as the partial credit algorithm was finely tuned to the ASSISTments platform and did not translate perfectly to Cognitive Tutor data. While the enhanced efficiency observed using partial credit was significant within ASSISTments, the gains were not statistically significant in Cognitive Tutor. Future work should examine the balance between enhancing the efficiency of measurement and the potential for overfitting.

An open question that remains following this work involves the potential link between observed findings and the amount of data considered. The partial credit algorithm used in the present work considered students' binary scores, the number of attempts required to solve an item, the number of hints required to solve an item, and a binary flag denoting whether the student was shown the answer through a "bottom out hint." Was partial credit more efficient than binary scoring in differentiating between groups because of the type of information used, or simply because additional information was added, giving a more robust view of student performance? When considering only three skill items, the method essentially transformed three binary data points into an aggregation of twelve pieces of data via partial credit scoring. Future research should extend the number of skill items examined and compare the effect of including various features. For instance, is the binary scoring of twelve skill items as effective as the partial credit scoring of just three skill items? What if six items are used but only binary accuracy and hint use are considered, resulting again in twelve pieces of data? If all findings were roughly equivalent, when would it be more appropriate to report partial credit scores rather than lengthen assignments?

The method utilized herein depicted how the resulting conclusions drawn from an educational dataset can vary based on data availability, aggregation, and manipulation. While the argument for more efficient group differentiation may not yet be a popular thesis for the learning analytics community, the benefits of partial credit scoring are likely much more broad. In experimental data, such as that originally investigated by Wang et al. (2016), the use of partial credit scoring can reduce costs associated with running randomized controlled trials (by reducing the number of subjects or items required to isolate

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

effective interventions). By examining non-experimental data, and specifically data sourced from mastery learning assignments, the present work has shown that partial credit scoring could also reduce the number of subjects or items required to build individualized learner models that attempt to predict student performance or proficiency. Future work should assess potential avenues for using group differentiation within learner models to better predict student mastery (i.e., groups that will reach mastery vs. those that will not). Additionally, the resampling approach presented herein could also be used for successful latent group differentiation, suggesting powerful implications for student modeling.

In closing, while it is not necessarily surprising that considering additional student data results in a more robust view of performance, thereby helping to distinguish groups of learners and strengthen predictive models, this observation speaks volumes about the flexibility of educational data. Researchers in learning analytics should remember to consider how accessible data can be "flexed" through cleaning and manipulation, and how this flexibility may impact observed outcomes and, ultimately, the field's understanding of best practices in education.

ACKNOWLEDGMENTS

We acknowledge funding from multiple NSF grants (ACI-1440753, DRL-1252297, DRL-1109483, DRL-1316736, and DRL-1031398), the U.S. Department of Education (IES R305A120125, R305C100024, and GAANN), the ONR, and the Gates Foundation. Thanks to S.O. & L.P.B.O.

REFERENCES

Attali, Y. (2011). Immediate feedback and opportunity to revise answers: Application of a graded response IRT model. *Applied Psychological Measurement, 35*(6), 472–479. https://dx.doi.org/10.1177/0146621610381755

Attali, Y., & Powers, D. (2008). *Effect of immediate feedback and revision on psychometric properties of open-ended GRE subject test items*. (ETS Research Report No. RR-08-21; GRE Research Report No. GREB-04-05). Princeton, NJ: Educational Testing Service.

Attali, Y., & Powers, D. (2010). Immediate feedback and opportunity to revise answers to open-ended questions. *Educational and Psychological Measurement, 70*, 22–35. https://dx.doi.org/10.1177/0013164409332231

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutor: Lessons learned. *The Journal of the Learning Sciences, 4*(2), 167–207. http://dx.doi.org/10.1207/s15327809jls0402_2

Carnegie Learning. (2016). Cognitive Tutor Software. *Carnegie Learning. Inc.* https://www.carnegielearning.com/learning-solutions/software/cognitive-tutor/

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*, 253–278. http://dx.doi.org/10.1007/BF01099821

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

Dasu, T., & Loh, J. M. (2012). Statistical distortion: Consequences of data cleaning. *Proceedings of the VLDB Endowment, 5*(11), 1674–1683. http://dx.doi.org/10.14778/2350229.2350279

Desmarais, M. C., & Baker, R. S. J. d. (2011). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, *22*(1–2), 9–38. http://dx.doi.org/10.1007/s11257-011-9106-8

Elgendy, N., & Elragal, A. (2014). Big data analytics: A literature review paper. In P. Perner (Ed.), *Advances in Data Mining: Applications and Theoretical Aspects* (ICDM 2014), Lecture Notes in Computer Science, vol. 8557 (pp. 214–227). Springer. http://dx.doi.org/10.1007/978-3-319-08976-8_16

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. *International Journal of Artificial Intelligence in Education, 24*(4), 470–497. http://dx.doi.org/10.1007/s40593-014-0024-x

Huff, D. (1954). *How to Lie with Statistics*. New York: W. W. Norton.

Janssen, M., van der Voort, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality. *Journal of Business Research*, 70, 338–345. https://dx.doi.org/10.1016/j.jbusres.2016.08.007

KDD Cup. (2010). Rules of the KDD Cup 2010: Educational data mining challenge. PSLC DataShop. https://pslcdatashop.web.cmu.edu/KDDCup/rules.jsp

Keller, K. L., & Staelin, R. (1987). Effects of quality and quantity of information on decision effectiveness. *Journal of Consumer Research*, *14*(2), 200–213. https://dx.doi.org/10.1086/209106

MATLAB version R.2013.a (2013). Natick, MA: The MathWorks. www.mathworks.com

Morgan, L. (2015). Flexibility is critical for big data analytics. *SD Times*. http://sdtimes.com/flexibility-is-critical-for-big-data-analytics/

NGACBP (National Governors Association Center for Best Practices) & CCSSO (Council of Chief State School Officers). (2010). *Common Core State Standards.* Washington, DC: Authors.

O'Reilly, C. A. (1982). Variations in decision makers' use of information sources: The impact of quality and accessibility of information. *The Academy of Management Journal*, 25(4), 756–771. https://dx.doi.org/10.1086/209106

Ostrow, K., Donnelly, C., & Heffernan, N. (2015). Optimizing partial credit algorithms to predict student performance. In O. C. Santos et al. (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining* (EDM2015), 26–29 June 2015, Madrid, Spain (pp. 404–407). International Educational Data Mining Society.

Ostrow, K., Donnelly, C., Adjei, S., & Heffernan, N. (2015). Improving student modeling through partial credit and problem difficulty. In Russell et al. (Eds.), Proceedings of the 2nd ACM Conference on Learning @ Scale (L@S 2015), 14–18 March 2015, Vancouver, BC, Canada (pp. 11–20). New York: ACM.

Ostrow, K., & Heffernan, C. (2016). *The ASSISTments TestBed resource guide.* www.assistmentstestbed.org

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

Ostrow, K., Heffernan, N., Heffernan, C., & Peterson, Z. (2015). Blocking vs. interleaving: Examining single-session effects within middle school math homework. In Conati et al. (Eds.), *Proceedings of the 17th International Conference on Artificial Intelligence in Education* (AIED '15), 22–26 June 2015, Madrid, Spain (pp. 388–347). Springer. http://dx.doi.org/10.1007/978-3-319-19773-9_34

Pardos, Z. A., & Heffernan, N. T. (2010). Modeling individualization in a Bayesian networks implementation of knowledge tracing. In Bra et al. (Eds.), *Proceedings of the 18th International Conference on User Modeling, Adaptation, and Personalization* (UMAP'10) 20–24 June 2010, Big Island, Hawaii, USA (pp. 255–266). Springer-Verlag Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-642-13470-8_24

Pardos, Z. A., & Heffernan, N. T. (2011). KT-IDEM: Introducing item difficulty to the knowledge tracing model. In J. A. Konstan, R. Conejo, J. L. Marzo, & N. Oliver (Eds.), *User Modeling, Adaption and Personalization*. (UMAP 2011). Lecture Notes in Computer Science, vol. 6787 (pp. 243–254). Springer, Berlin, Heidelberg. http://dx.doi.org/10.1007/978-3-642-22362-4_21

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, *14*(2), 249–255. http://dx.doi.org/10.3758/BF03194060

Stamper, J. C., Koedinger, K. R., Baker, R. S. J. d., Skogsholm, A., Leber, B., Demi, S., Yu, S., & Spencer, D. (2011). DataShop: A data repository and analysis service for the learning science community. In Biswas et al. (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (AIED '11), 28 June–2 July 2011, Auckland, New Zealand (pp. 557–559). Springer. http://dx.doi.org/10.1007/978-3-642-21869-9_129

VanLehn, K., Lynch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., & Wintersgill, M. (2005). The Andes physics tutoring system: Lessons learned. *International Journal of Artificial Intelligence in Education*, *15*(3), 147–204.

Wang, Y. (2016). Data and code for How Flexible Is Your Data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. http://tiny.cc/JLA_ShapeOfLearning

Wang, Y., & Heffernan, N. T. (2011). The "Assistance" model: Leveraging how many hints and attempts a student needs. *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference* (FLAIRS-11), 18–20 May 2011, Palm Beach, FL, USA (pp. 549–554). Menlo Park, CA: The AAAI Press.

Wang, Y., & Heffernan, N. (2013). Extending knowledge tracing to allow partial credit: Using continuous versus binary nodes. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education* (AIED '13), 9–13 July 2013, Memphis, TN, USA (pp. 181–188). Springer. http://dx.doi.org/10.1007/978-3-642-39112-5_19

Wang, Y., Ostrow, K., Beck, J., & Heffernan, N. (2016). Enhancing the efficiency and reliability of group differentiation through partial credit. In Gašević et al. (Eds.), *Proceedings of the 6th International Conference on Learning Analytics and Knowledge* (LAK '16), 25–29 April 2016, Edinburgh, UK (pp. 454–458). New York: ACM. http://dx.doi.org/10.1145/2883851.2883910

(2017). How flexible is your data? A comparative analysis of scoring methodologies across learning platforms in the context of group differentiation. *Journal of Learning Analytics*, *4*(2), 91–112. http://dx.doi.org/10.18608/jla.2017.42.9

Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized Bayesian knowledge tracing models. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education* (AIED '13), 9–13 July 2013, Memphis, TN, USA (pp. 171–180). Springer. http://dx.doi.org/10.1007/978-3-642-39112-5_18