ESTABLISHING STATISTICAL SIGNIFICANCE FOR COMPARISONS USING PATTERN-BASED ITEMS: CHANGE AT SCALE

Walter M. Stroup
University of Massachusetts, Dartmouth
wstroup@umassd.edu

Anthony Petrosino
Southern Methodist University
apetrosino@smu.edu

Corey Brady
Southern Methodist University
corey.brady@smu.edu

Karen Duseau
Bridgewater State University
kduseau@bridgew.edu

Tests of statistical significance often play a decisive role in establishing the empirical warrant of evidence-based research in education. The results from pattern-based assessment items, as introduced in this paper, are categorical and multimodal and do not immediately support the use of measures of central tendency as typically related to interpretations of measures of statistical significance. Responses from the duplicate implementation of selected pattern-based items (PBIs) in successive grades (3-8) as part of the statewide Interim Assessment Program in Texas are used to illustrate how non-parametric methods can be used to establish statistically significant comparisons of student results. Not all the repeat-item results improved across years.

Keywords: Assessment, Policy, Research Methods, Systemic Change

During an assessment window from August 31, 2018 through March 31, 2019 a series of non-dichotomous, or "pattern-based," items, focused primarily on mathematics but also including reading items, were implemented with more than 400,000 students in grades three through eight on a statewide assessment program. The items were part of an opt-in Interim Assessment Program made available to districts by the Texas Education Agency (TEA). A central question addressed in a report to TEA (Stroup, 2020) was whether the greater information density of pattern-based items (PBIs) would provide significant, actionable insights into student learning outcomes. A central focus of this paper, however, is to use results from the statewide implementation to outline how measures of statistical significance can be established for PBIs.

Results from pattern-based assessment items, as presented below, are categorical and multimodal and do not immediately support the use of measures of central tendency, as typically related to interpretations of measures of statistical significance. As will be illustrated with across-grade results, using either a Fisher's Exact Test or a Pearson's Chi-squared Test we can evaluate the statistical significance of differences in the multimodal patterns of responses. Then, once a level of statistical significance is arrived at, projections onto axes of relative performance – e.g., assigning "partial credit" or using exact full credit scoring – can be used for comparisons of achievement-related outcomes for groups of students.

Pattern-based items were developed by teachers and classroom-focused researchers to provide colleagues -- as part of supporting ongoing instruction -- with a more detailed "picture" of students' understandings (Stroup, 2020). Consistent with this year's conference theme, classroom assessments, support by shared display capabilities, immediate forms of feedback, and suggestions for further inquiry, should center on engaging *all* learners. Accordingly, pattern-based approaches to assessment are meant to support students' full participation in ongoing classroom-based teaching and learning. However, educator reports of the effectiveness of pattern-based approaches in supporting ongoing instruction may not be enough. We live at a time when phrases like “evidence-based practice” are assumed to denote a highly constrained set of methodological commitments. Absent ways of engaging in forms of hypothesis testing, centered on measures of statistical significance, it may be too facile for critics of understandings-focused classroom assessment (*cf.*, Bennett, 2011) to attempt to limit the reach of educator-driven, pattern-based approaches to assessment. In response, the focus of this paper is on the use of PBIs to support making empirically warranted claims that might, for example, be plausibly related to differences in instruction between individual classrooms, schools, districts, or states and/or to differences in instruction organized in terms of temporal change, or “growth”, in student achievement.

What is a Pattern-Based Item?

Pattern-based items are developed to provide a significant, and fully scalable, alternative to current, dichotomously scored and analyzed, items. While PBIs may appear similar to standard multiselect multiple-choice items, they should be seen as fundamentally distinct, both in terms of how they are developed and in terms of how they are analyzed.

For an individual assessment question with four responses – A, B, C, or D – both multiselect and pattern-based items allow the students to select, or endorse, more than one response (e.g., “A and C” or “B, C, and D”). While Rasch-based partial credit models have been explored since the early 1980’s (Masters, 1982), it remains the case that standard multiselect items have only one combination, the “exact match”, be scored as correct (“1”). All other combinations are typically scored as incorrect (“0”). While these “legacy” multiselect multiple-choice items might be viewed as an improvement over the more commonly used single-select multiple-choice items, the subsequent (non-polytomous) analyses, for both single-select and multiselect items, continue to be based, in most cases, on only two “states”, either 1 or 0.

In contrast, the pattern-based items appearing on the Interim Assessments in Texas were developed to work with the full combinatoric space of student responses. This means that for the PBIs discussed below, there are sixteen possible selection combinations of the four available responses (or fifteen, if “no response” is not included). Rather than continue the current practice of reducing these states to only two states (i.e., correct/incorrect), the greater information density of pattern-based items (sixteen states versus two states) is meant to allow for more detailed, actionable insights related to student learning outcomes. More information about their students supports teachers in improving student outcomes. Pattern-based assessments can be shorter and more informative than legacy assessments in ways that are meant to directly support better instruction. Although not discussed in this paper, pattern-based items or tasks are not limited to interactions where the student selects among given responses (*cf.* Stroup [1996], Stroup & Wilensky [2000]).

Differences in the patterns in responses of student groups to PBIs can be analyzed across scale and also over time, allowing for more detailed longitudinal evaluation of teaching and learning at the class, school, district, and state levels. Based on these uses of PBIs, our task is to provide a framework for establishing measures of item-level statistical significance as illustrated using observed differences in student results across grade levels. The framework is intended to serve as a widely-applicable, principled, approach to the determination of significance for comparisons using pattern-based items and tasks.

Implementation

The students participating in the state-sponsored Interim Assessment Program were required to complete a series of single-select multiple-choice State of Texas Assessments of Academic Readiness (STAAR) items prior to then having the option to complete the pattern-based items. The directions stated: “The next set of questions is optional. These questions will not be counted as part of your score.” Moreover, due to terms agreed to by the Texas Education Agency (TEA) and the schools participating in the Interim Assessment Program, no datasets containing student-, school-, or district-identifiable data would be provided by the vendor to the Agency.

Consistent with these terms, we received from the TEA two, large, fully anonymized, datasets containing only each student’s selections for the pattern-based items. As a result, the data cannot be used to provide an account of the makeup of the students who participated in the overall Interim Assessment Program or of those who then elected to complete the optional items. Comparisons between results for dichotomous legacy items and results for PBIs on the Interim Assessments are also precluded. To be able to illustrate how PBIs support comparisons across scale – e.g., comparing individual classroom results or school-wide results with statewide results – in the Report to TEA (Stroup, 2020) we augmented the datasets from the Interim Formative Assessment Program with datasets from a December 2018 pilot implementation of the same items in two elementary schools in central Texas.

The average completion rate for the last item was more than 98% of the average completion rate for the first item. If there were difficulties with the implementation of items at scale, then one would expect much less consistency in the levels of participation. No issues with the statewide implementation were reported.

Comparing changes across grades

Even absent information about student, school, district, or temporal implementation (within the seven-month window) in the Interim Assessment Program dataset provided by the Texas Education Agency, the sensitivity of pattern-based items to one kind of overall student growth can be assessed by comparing results for items deployed across grade levels. The pattern-based item shown in Figure 1 assesses student understanding of equivalent ratios in relation to adding drops of blue food coloring to water to create a blue solution. As also shown in Figure 1, the fifteen letter combinations (“no response” is not included) can be shaded and sorted from lowest partial credit score (light shading) on the left, up to the full credit score (dark shading) on the right.

Partial credit, in this context, is assigned in terms of the degree of match with the full-credit response. With PBIs, not selecting an incorrect response is typically assigned the same partial credit as selecting a correct response. For this item, then, full credit is assigned to selecting B, C, and D and not selecting A. The histograms of results shown in Figure 1 are sorted from zero partial credit for only selecting the one incorrect response, A, up to the full-credit response of BCD (where not selecting A is implicit in A not appearing in labelling of the histogram bin). The assigning of credit in this way is treated as a projection from the combinatoric space of the students’ actual selected responses onto a single axis of relative performance.

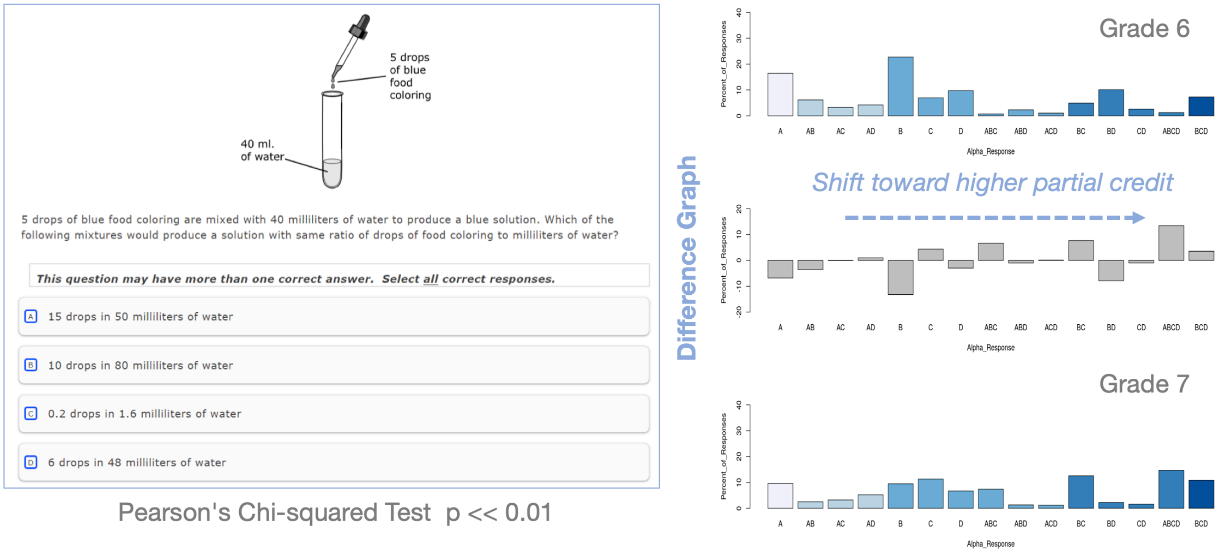
Of course, other partial credit projections are also available. The most widely deployed among these alternatives might be the partial credit models used by most learning management systems (c*f.* Jones, [2016]). Assuming that whatever model is deployed is consistent in how the multimodal results from pattern-based items are projected onto a partial credit axis, if statistical significance is established at the level required, then comparisons of partial credit results may be used to represent relative changes in student outcomes or as an overall measure of effect.

A key curricular transition in moving from late elementary grades to middle school grades is extending emergent multiplicative forms of reasoning about fractions to comparisons of ratios and proportions. To be able to assess the development of students' ratio and proportional reasoning as they enter, and then begin to move through, middle school, this pattern-based item was included on both the grade 6 (N=58,947) and grade 7 (N=46,483) Interim Assessments. To further situate how it is that the use of the same item across grade levels can be informative at both levels, we begin with a brief discussion of some of the state's related curriculum standards.

A Grade 6 mathematics Texas Essential Knowledge and Skills (TEKS) standard requires students to be able to "apply qualitative and quantitative reasoning to solve prediction and comparison of real-world problems involving ratios and rates" (6.b.4.B) and "give examples of ratios as multiplicative comparisons of two quantities describing the same attribute" (6.b.4.C). To satisfy the Grade 7 TEKS, students must be able to "solve problems involving ratios" framed more explicitly in terms of proportional reasoning (7.b.4.D). With the item shown in Figure 1, students are assessed on their ability to use "multiplicative comparisons of two quantities" (drops of food coloring and amount of water) in a real-world context to describe the "same attribute" of what would be the blueness of the resultant solutions.

At both grade levels, students will often attempt to extend additive forms of reasoning to a task that requires multiplicative comparisons. Response A is intended to assess whether students are attempting to reason additively – adding 10 to both the number of drops and to the amount of water – about a task requiring multiplicative reasoning. The difference graph shown in Figure 1 reflects a 6.8% decrease in this pattern of reasoning: moving from 16.4% of the grade 6 students selecting only A to 9.6% of the grade 7 students selecting only A.

In the context of the design of pattern-based items, this decrease illustrates the significance of students' *not* selecting a response in contributing to an overall assessment of the depth of their understanding. In contrast to this additive, incorrect, answer, response B correctly multiplies the 5 drops of food coloring by 2 and the 40 milliliters of water by 2. Response D requires students to simplify the given ratio of 5 drops of food coloring to 40 milliliters of water to the equivalent "unit ratio" of 1 drop of food coloring to 8 milliliters of water, and then correctly multiply each quantity by 6. Response C also requires this simplification of the original ratio and then correctly multiplying each quantity by 0.2. Increasing depth of understanding is assessed in this pattern-based item in moving from not selecting A, to selecting B, to selecting D and then, at the highest level of understanding, selecting C.



5 drops of blue food coloring are mixed with 40 milliliters of water to produce a blue solution. Which of the following mixtures would produce a solution with same ratio of drops of food coloring to milliliters of water?

This question may have more than one correct answer. Select all correct responses.

- A 15 drops in 50 milliliters of water
- B 10 drops in 80 milliliters of water
- C 0.2 drops in 1.6 milliliters of water
- D 6 drops in 48 milliliters of water

Figure 1. Analysis of differences in item results at scale.

The graph in the middle of Figure 1 shows all the relative changes in the percentages for each of the combinations of responses. Although the shifts can appear complex across the various combinations, there remains general movement from lower partial credit responses to higher partial credit responses. The respective percentages receiving full credit are 7% for grade 6 and 11% for grade 7 and the respective average partial credit scores are 47% and 56%.

Table 1 shows the contingency table for the 15 combinations of responses for each grade level.

Table 1: Observed frequencies of responses for IAP implementations of the same pattern-based mathematics item in grades 6 & 7

Responses	A	AB	AC	AD	B	C	D	ABC	ABD	ACD	BC	BD	CD	ABCD	BCD
Students															
IAP Grade 6	9625	3589	1919	2470	13297	4072	5684	427	1350	615	2871	5919	1515	736	4275
IAP Grade 7	4466	1164	1495	2418	4409	5255	3113	3431	601	553	5830	1033	738	6828	5054

The selection of the respective combinations and the result across grades (no students took both grade-level assessments) can be reasonably assumed to be independent. Then, for this comparison, the substantial number of students in all the cells at each grade level supports the use of a non-parametric Pearson's Chi-squared test with simulated p-value to evaluate statistical significance. With the p-value being 0.0005, these results are highly statistically significant.

The very low p-value supports the subsequent claim that positive differences in the projected outcomes on this item are of high statistical significance. Specifically, the positive differences in full credit and partial credit results for grade 7 compared to grade 6 can be represented as statistically highly significant. This ability to show relative improvement over time from grade 6 to grade 7, using any one of a number of possible projected metrics, illustrates the ability of pattern-based items to be used to assess changes in scores at scale. We would even argue that this logical transfer of statistical significance would extend to the 6.8% decrease in moving from 16.4% of the grade 6 students selecting only A to 9.6% of the grade 7 students selecting only A. Attending to students' not choosing certain responses is important to the development, analysis, and pedagogical utility of PBIs.

Extending the Utility of Reporting Effect in Terms of Partial Credit Projections

The evaluation of variations in outcomes for specific within-year treatments – e.g., a new curriculum – or comparisons of longitudinal growth for given groups of students would require, as would be the case with existing legacy items, a more sophisticated research design than was possible within the structure of the Interim Assessment Program in Texas. We can, however, illustrate how the evaluation of statistical significance, when combined with partial credit projections, can provide an account of overall levels of standards-specific achievement across years.

Five PBIs were repeated across grade levels. Figure 2 depicts the partial credit results for each of these items by grade level. Each point represents between forty and sixty thousand students in each grade. Scores for three of the items increased in ways suggesting improvement in the levels of standard-specific attainment across the respective grades.

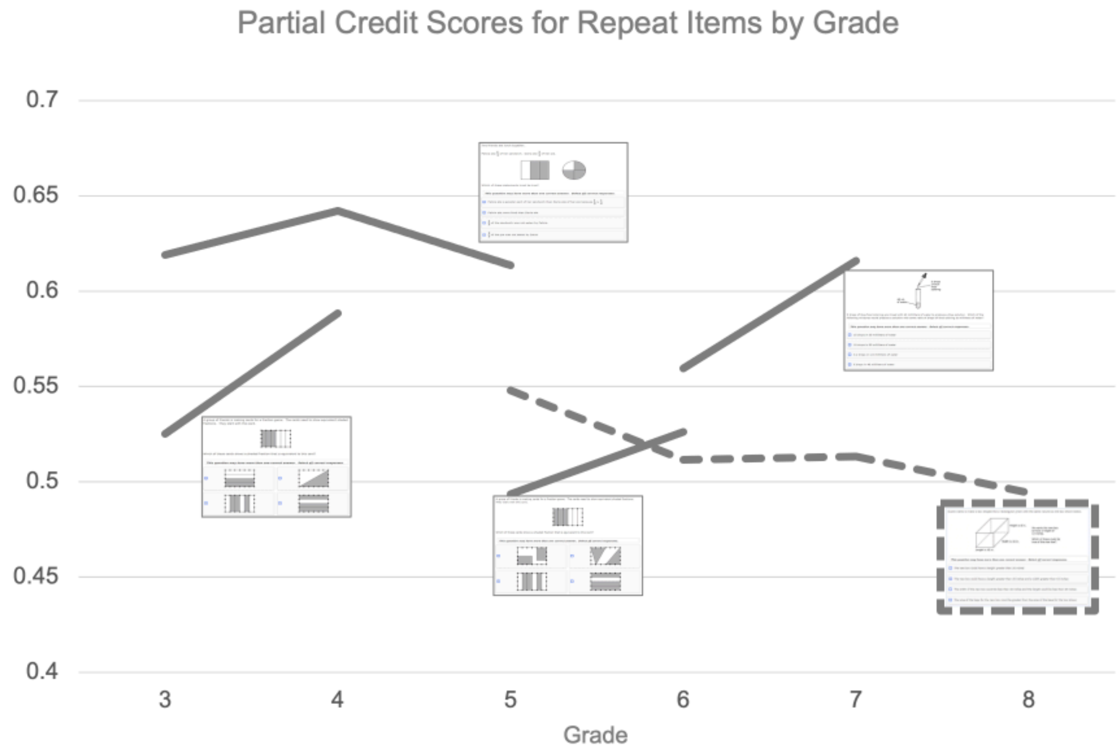
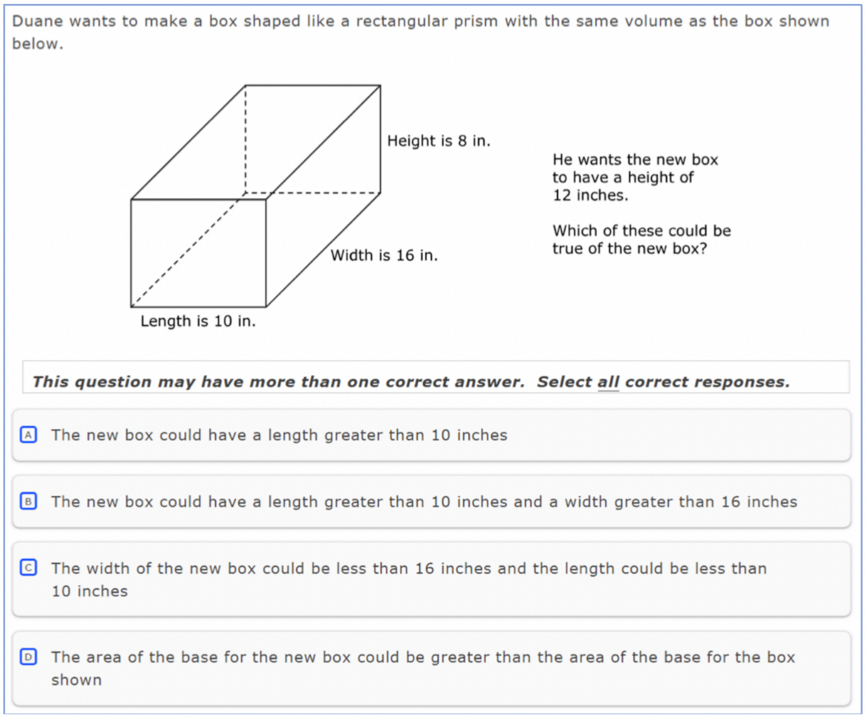


Figure 2. Partial credit results for repeated pattern-based items deployed across the grade levels shown.

Results for one repeated fractional-comparison question went up and then down. Of greater concern, however, is how the partial credit score for a volume item (the dashed line in Fig. 2) went down from grade five to grade eight. It considers a new box with the same volume and one known dimension.

Volume appears in the state curriculum standards for each grade assessed. As shown in Figure 3, the question starts with a given rectangular box and considers a new box with the same volume and one known dimension.



Duane wants to make a box shaped like a rectangular prism with the same volume as the box shown below.

He wants the new box to have a height of 12 inches.

Which of these could be true of the new box?

This question may have more than one correct answer. Select all correct responses.

A The new box could have a length greater than 10 inches

B The new box could have a length greater than 10 inches and a width greater than 16 inches

C The width of the new box could be less than 16 inches and the length could be less than 10 inches

D The area of the base for the new box could be greater than the area of the base for the box shown

Figure 3. Volume item for which partial credit scores trended downward.

The student is asked about various possible changes to the box that could keep the volume the same (Stroup, 2020). When we've asked educators what might explain the decrease in scores for the volume item, some have suggested there may be an over-reliance in upper grades on formalisms to give "the answer." Students may come to rely on "the formula" (Volume = length*width*height) in a way that leaves them less able to think conceptually about how a specified change in one dimension could be compensated for with possible changes in the remaining dimensions.

Independent of what the explanations or the possible pedagogical responses might be for specific scores on a given item going up or down, the larger point, illustrated by the graph shown in Figure 2, is that the use of partial credit projections is a relatively transparent way to interpret results having statistical significance. Partial credit scores going up, or down, by some percent is likely to be more widely understood than, for example, using Cramer's V as a way to discuss effect size. The generation of a Cramer's V value between 0 and 1 can be considered relatively opaque, certainly in comparison to calculating partial credit. Moreover, the values used to distinguish between "weak" (0.1-0.3), "medium" (0.4 to 0.5), and "strong" (>0.5) associations using Cramer's V are simply conventions that have evolved to become accepted standards in practice. As is illustrated in Figure 2 and in an earlier example, we would suggest using comparisons of partial credit results as a preferred way of characterizing relative differences in student outcomes.

Summary and Conclusion

In order for assessments based on the use of pattern-based approaches to be more widely deployed as a significant, and fully scalable, alternative to current assessments using dichotomously scored and analyzed items, the results from PBIs must be useful to educators' efforts to engage and support all learners. This, however, may not be enough. Results from assessments are frequently used to make statistically warranted claims about comparisons in outcomes. There need to be methods for evaluating the statistical significance of differences in PBI results for groups of students. The utility of PBIs for educators is addressed more directly elsewhere. This paper, instead, has focused on the issue of evaluating statistical significance.

Acknowledgments

This work was supported by funding from the National Science Foundation under Grant No. DRL-16152 (Group-based Cloud Computing for STEM Education). The views expressed are solely those of the authors.

References

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5-25. https://doi.org/10.1080/0969594X.2010.513678

Jones, J. (2016, April 21). *Understanding multiple answers questions*. Instructure. https://community.canvaslms.com/t5/Higher-Ed-Canvas-Users/Understanding-Multiple-Answers-Questions/ba-p/263903

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174. https://doi.org/10.1007/BF02296272

Stroup, W. (1996). *Embodying a nominalist constructivism: Making graphical sense of learning the calculus of how much and how fast* (Unpublished doctoral dissertation). Harvard University, Cambridge, MA.

Stroup, W. M., & Wilensky, U. (2000). Assessing learning as emergent phenomena: Moving constructivist statistics beyond the bell curve. In A. E. Kelly & R. A. Lesh (Eds.), *Handbook of research in mathematics and science education*. Mahwah, NJ: Lawrence Erlbaum Associates.

Stroup, W. M. (2020). *Interim Assessment Pattern-Based Item Report* (Unpublished report). Submitted to the Texas Commissioner of Education on December 12, 2019 and approved for release on June 12, 2020. Retrieved from https://drive.google.com/file/d/1krNGNe97YIe1GVCo0g-vz9zEnX0r09Ie/view?usp=sharing