Measuring the Power of Learning.®

Research Report
ETS RR–17-32

Variability in Percentage Above Cut Scores Due to Discreteness in Score Scale

Ying Lu

June 2017

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

RESEARCH REPORT

Variability in Percentage Above Cut Scores Due to Discreteness in Score Scale

Ying Lu

Educational Testing Service, Princeton, NJ

For standard- or criterion-based assessments, the use of cut scores to indicate mastery, nonmastery, or different levels of skill mastery is very common. As part of performance summary, it is of interest to examine the percentage of examinees at or above the cut scores (PAC) and how PAC evolves across administrations. This paper shows that discreteness in score scales can affect the PAC statistics considerably, especially when the test is short and when there are high examinee density and low item density near the cut score on the reporting score scale. This paper also includes recommendations on how to adjust the PAC statistics when they are used in trend analyses.

Keywords Cut scores; criterion-based assessment; accountability; growth; percent proficient; percentage above cut scores

doi:10.1002/ets2.12158

For standard- or criterion-based assessments, the use of cut scores to indicate mastery, nonmastery, or different levels of skill mastery is very common. Cut scores have an important meaning in the score scale and represent the consensus reached by content experts over definitions of achievement levels. Naturally it is of interest to examine the percentage of examinees at or above the cut scores (PAC) as part of performance summary and the trend of how this statistic evolves across administrations.

In recent years, K – 12 large-scale educational assessments, with the goal of having all students achieve at least the proficient level, have drawn intense attention to PAC statistics. The percentage of examinees scoring at or above the proficient level is the cornerstone statistic of the accountability system and has been used for adequate yearly progress determination for schools. This proficiency framework has also led to the use of the change in the percentage of proficient students across years as one of the most popular trend statistics. Positive trends are associated with interpretations such as teacher effectiveness and educational improvement. Almost all states report state-level percentage-proficient statistics for their K – 12 test assessments and the yearly change of the percentage-proficient statistics in a trend analysis report. As an example, Figure 1 shows the yearly change in state-level percentage proficient for reading and mathematics from 2002 to 2010 for the Florida Comprehensive Assessment Test. While the plots show a general upward trend in performance as reflected by positive yearly change in percentage-proficient statistics, they also show considerable variation in percentage-proficient statistics for reading and some moderate variation for mathematics. Note that district and school level results may show more variations due to smaller sample sizes based on which the proficiency rates are calculated.

PAC trend statistics are not only summarized over a single testing program, they are also compared across different testing programs. For example, studies (Ho, 2007; Schafer, Liu, & Wang, 2007) have compared state test-score trends to the National Assessment of Educational Progress (NAEP) score trends in the form of changes in PAC statistics. Discussion of the results of these comparison studies is beyond the scope of this paper, but these studies showed the popularity of the use of PAC statistics in trend analyses and how closely they are scrutinized in making important educational policy and reform decisions.

The use of PAC statistics for trend analysis and cross-test comparison is not without criticism, and alternatives have been suggested. Holland (2002) as well as Ho (2008) criticized the universal use of percentage-proficient statistics in large-scale educational assessments by showing that proficiency-based trends and gap trends are subject to changes in magnitude and sign under alternative proficiency cut scores. As an alternative to PAC statistics, the PP plot, also called the proportion-proportion or percentage-probability plot (Wilk & Gnanadesikan, 1968), has been proposed by Haertel,

Corresponding author: Y. Lu, Email: ylu@ets.org

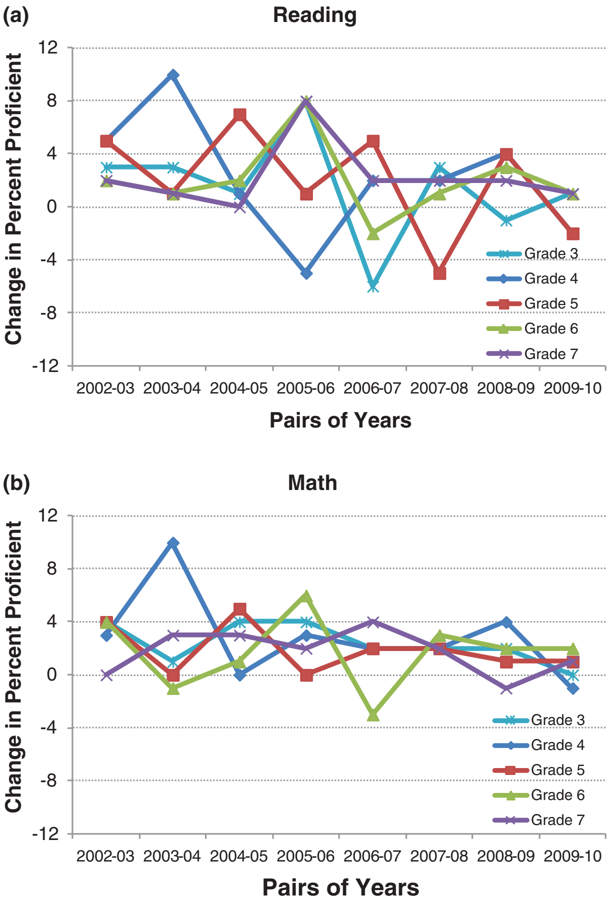

Figure 1 Change in state-level percentage proficient for reading and mathematics from 2002 to 2010 for Florida's Comprehensive Assessment Test, Grades 3 to 7 (Florida Department of Education, n.d.).

Thrash, and Wiley (1978), Spencer (1983), and Livingston (2006) for the comparison of test-score distributions. Further, Ho (2009) suggested a nonparametric framework for comparing trends and gaps across tests. Nevertheless, the use of PP and other PAC statistics in trend analyses is popular, given the important meanings represented by the cut scores.

When conducting trend analysis using PAC statistics associated with a particular *pass* or *proficiency* criterion, it is important to be aware of various factors that contribute to the annual changes in the percentage at performance level. Some factors (e.g., instructional effectiveness) are intended to be captured in the yearly change statistics. Others, such as random variation in student intake, measurement error, equating error, and different tests taken, are irrelevant and may lead to misleading classification rates and year-to-year trends. Researchers have conducted various studies to evaluate the degree that the irrelevant factors affect classification rates and hence accuracy in the performance trend. Arce-Ferrer, Frisbie, and Kolen (2002), Linn and Haug (2002), Kane and Staiger (2002), and Yen (1997) pointed to sampling error as a significant factor accounting for variability in achievement-level percentages. Betebenner, Shang, Xiang, Zhao, and Yue (2008) quantified the extent to which measurement error produced bias and increased error variability for percentage at performance-level statistics. Koretz (2005) observed that teaching to the test and other inappropriate test preparation may substantially inflate the gains on high-stakes tests.

One source of unreliability that is rarely scrutinized by researchers is the discreteness in score scales. For criterion-referenced tests, cut scores are commonly established in the first year of a testing program to report student scores. To distinguish them from the empirical cut scores used in later administrations, these cut scores are referred to in this paper as *theoretical cut scores*. In later administrations of the same testing program, different test forms are administered and equating is conducted to take into account form-to-form difference in difficulty. A student is regarded as meeting the standard if the student's equated score is at or higher than the theoretical cut score. For tests that report scale scores based on a raw-to-scale score conversion table, the obtainable scale score that is at or just above the theoretical cut score is

Table 1 2013 and 2014 Raw-to-Scale Score Conversion Table for a Hypothetical Test

	2013		2014	
Raw score	Conversion	% at and above	Conversion	% at and above
..	..	..	..	
43	455	52	459	48
42	450	56	454	54
41	446	61	449	58
40	442	66	445	64
..	..	..	..	

Note. Underlining indicates cut scores.

referred to as *empirical cut score* in this paper. Empirical cut scores are always higher than or equal to the theoretical cut score except for the situation when rounded scale scores are used for reporting, and empirical cut scores can be within half point below the theoretical cut scores. When empirical cut scores differ systematically from test form to test form, the PAC statistics are affected accordingly. With the same theoretical cut score, the PAC statistics tend to be lower when the empirical cut scores are higher and vice versa. Although tests that adopt item response theory (IRT) model pattern scoring may be less affected by the issue, the majority of the K–12 state assessments, large-scale assessments, and licensure tests report scale scores based on a raw-to-scale score conversion table. The difference in the theoretical and empirical cut scores is mostly due to discreteness in score scales, test form variability, and the fact that it is unrealistic to administer a test with an infinite number of items to the examinees. It is an issue that cannot be resolved by equating.

To illustrate what it means to have variant empirical cut scores from one administration to another, Table 1 shows part of the 2013 and 2014 raw-to-scale score conversion table for a hypothetical test that has 60 items. Suppose that the proficiency scale cut score for this test was established to be 450 (i.e., the theoretical cut score) based on agreement by content experts and policy makers. The table shows that the 2013 and 2014 forms had the same raw cut scores, with 42 being the lowest raw score associated with a scale score of 450 or above. However, the minimum obtainable scale score required to be proficient (i.e., the empirical cut score) was 450 in 2013 and 454 in 2014, due to difference in test form difficulty. Therefore, although the theoretical cut score of 450 did not change, students had to achieve a higher scale score to be proficient in 2014 than in 2013. This makes it difficult to capture the performance trend accurately as the 2014 sample might show less growth than expected due to the boost in the 2014 empirical cut score. Theoretically, the impact may be alleviated by adding sufficient number of items to the 2014 test form so that there is a raw score that corresponds more closely to a theoretical cut score of 450. This, however, is not viable due to issues such as fatigue and time limit.

The purpose of this paper is to examine the extent to which PAC statistics are affected by the issue of discreteness in score scales and further to provide recommendations on how PAC statistics should be used in trend analyses.

Method

Theoretically quantifying the variation in PAC statistics due to discreteness in score scales is difficult and may involve numerous assumptions regarding score distributions and test properties that may not necessarily be valid. An ad hoc procedure would be to require the examinees to take numerous test forms with the same number of items, built under the same blueprint and of similar difficulty as the test that the examinees actually took. With the same theoretical cut score, the different test forms will yield different empirical cut scores and hence different values for the PAC statistics. Realistically this procedure would be difficult to implement. If student true ability and response models are known, simulations may be conducted to generate student responses to hypothetical test forms, and PAC statistics can be calculated on the hypothetical test forms. While these kinds of simulations are possible in research studies, more helpful to practitioners would be a method that does not rely on true examinees and true response models that are not known in real testing situations.

To evaluate PAC variation, the study used the delete-one jackknife method for resampling and creating a pool of test forms that were similar in nature to the original test form the students have taken. The jackknife method was originally suggested as a method of reducing bias (Quenouille, 1949, 1956) and was later used to estimate variances and calculate confidence intervals (Tukey, 1958). In this study, the jackknife bias and variance calculation method is not applied in estimating PAC, as the function of PAC is determined using observed item responses and is not capable of Taylor expansion. Only the delete-one jackknife resampling procedure was used in the study.

Specifically, if the students have taken a test form that consists of n items, the delete-one jackknife procedure creates a pool of n new unique forms with $n-1$ items on each of the new forms (i.e., the first form has all items from the original form except for Item 1; the second form has all items from the original form except for Item 2). The n new forms are only one item short of the original test form and can be regarded as roughly equivalent to the original form and to each other in terms of content and form difficulty. Equating is then conducted based on each of the n new forms to produce n new raw-to-scale score conversions for examinee scoring. Students will be assigned n new scale scores based on n new forms and will also receive n additional performance classifications. The variation of the performance classifications across test forms that are considered semi-equivalent in construct will then be examined.

To examine the variation of PAC statistics under various test conditions, simulations were designed and conducted in this study. Note that the major purpose of using simulations was to control varying factors that were expected to affect the estimation of PAC statistics. These factors include test length, cut-score location, examinee sample size, and so forth. The use of simulations also ruled out the variation of PAC statistics due to item calibration errors or model data misfit when IRT is used for parameter estimation and for generating raw-to-scale score conversion tables. The procedure itself, however, can be applied to any operational form relatively easily without conducting any simulations.

The simulations treated test length and cut-score location as major factors for control. Hypothetical test forms of 20, 40, 60, and 80 dichotomous items were created. These test forms were assumed to have a score scale of 100 to 700 with 100 being the lowest obtainable scale score and 700 being the highest obtainable scale score. Cut scores were defined to be 400, 450, 500, and 550. Examinee sample size took the values of 100,000, 1,000, and 200. The large size of 100,000 was used to rule out the factor of sample variation, so that the impact of discreteness in score scale can be examined with minimum confounding factors. The sample sizes of 1,000 and 200 were used to approximate district and school size when calculating a grade-level PAC at the local level.

Three sets of simulations were conducted. The first set of simulations attempted to emulate a realistic operational testing condition. The second and third sets of simulations intended to disentangle the interactions among cut-score locations, item density, and examinee density at the cut-score area. In all the three sets of simulations, an IRT two-parameter logistic (2PL) model was used for student response generation with the discrimination parameter a_i randomly selected from a log-normal distribution having a mean of 1.0 and standard deviation of 0.2. In the first set of simulations, both item difficulty parameter b_i and examinee true abilities were generated from a standard normal distribution. This practice is consistent with common testing situations when item difficulties and examinee abilities are well matched and both exhibit somewhat normal distributions. In the second set of simulations, item difficulty parameters were generated from a uniform distribution from -2 to 2 and examinee abilities were generated from a standard normal distribution. In the third set of simulations, item difficulty parameters were generated from a standard normal distribution and examinee abilities were generated from a uniform distribution from -2 to 2.

The raw score to ability conversion table was generated based on the true item parameters for each generated test form using IRT true score equating (Kolen & Brennan, 2004; Stocking & Lord, 1983). Note that true item parameters instead of calibrated item parameters were used in order to rule out IRT calibration error and IRT model data misfit as additional confounding factors.

Linear transformation was applied to examinee abilities on the IRT scale to produce the raw-to-scale score conversion table. A hypothetical slope of 100 and intercept of 400 were applied in the transformation. Note that the choice of slope and intercept was totally arbitrary and does not affect the results of this study. These choices were used to make the test data in this study look realistic. When examinee abilities were generated from a standard normal distribution, the transformed abilities have a mean of 400 and a standard deviation of 100. Under conditions when examinee abilities were generated from uniform distribution of -2 to 2, the transformed abilities had a uniform distribution from 200 to 600. Gauging cut scores of 400, 450, 500, and 550 against the transformed ability distributions provided information on examinee density at cut scores of interest. Once raw-to-scale score conversions were generated for each test form as well as its jackknife replicates, examinees were scored and classified based on the cut scores. In all simulations in this study, rounded scale scores were assigned to students, meaning that a student obtaining an actual score of 399.5 or more was considered at or above level when the cut score was 400. The rounding procedure was used because this method is consistent with the practices of most testing programs reporting scale scores.

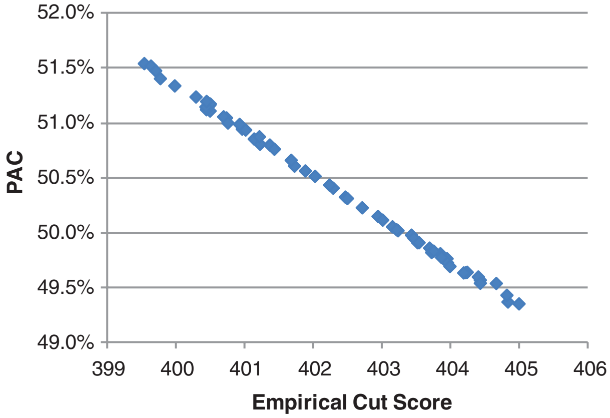

Figure 2 Percentage of examinees at or above the cut scores (PAC) and empirical cut score (test length = 60, theoretical cut score = 400, sample size = 100,000, *ability* ~ $N(0,1)$, $b_i \sim N(0,1)$).

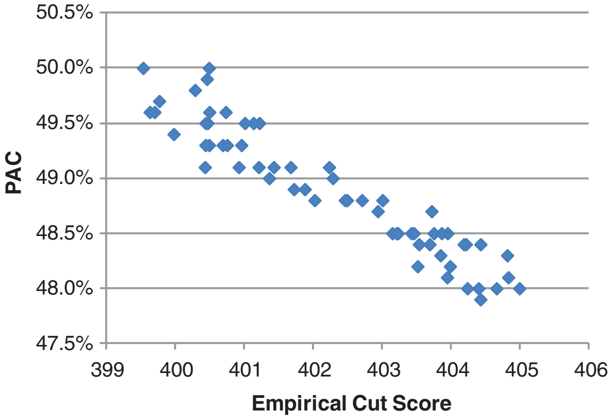

Figure 3 Percentage of examinees at or above the cut scores (PAC) and empirical cut score (test length = 60, theoretical cut score = 400, sample size = 1,000, *ability* ~ $N(0,1)$, $b_i \sim N(0,1)$).

Results: PAC Variation

The resulting PAC statistics were plotted against the empirical cut scores based on the original generated test form and all the jackknife replicates under all simulated conditions. Even though the theoretical cut scores remain unchanged, different test forms lead to different empirical cut scores. Without exception, the plots follow a linear pattern between PAC values and the empirical cut scores. As expected, the higher the empirical cut score is, the lower the PAC is.

For illustration purpose, Figures 2–4 show scatter plots from the first set of simulations with a 60-item test and a cut score of 400 under different examinee sample sizes. Given a sample size of 100,000 that rules out nearly all effect of examinee sample variation, Figure 2 shows a clear straight line with minimum fluctuation or scatter. Figures 3 and 4 show more scatter due to sample variation. Figure 2 shows that for the theoretical cut score of 400, the empirical cut score may range from 399.5 to about 405 depending on the specific test form an examinee is given. The PAC statistic associated with each empirical cut score ranges from 49.4 to 51.6. This range indicates that under the specific simulation condition, as much as (51.6 − 49.4) = 2.2 in PAC variation can be attributed to the different forms taken and different empirical cut scores applied even after equating takes into account form-to-form difference. Figures 3 and 4 show similar PAC variation, but with more scatter.

The PAC range associated with the same theoretical cut scores given different test forms, called *PAC variation range* in this paper, is used to summarize PAC variation under different conditions and is defined as follows:

$$\begin{aligned}\text{PAC variation range} = {} & \max(\text{PAC values based on different forms of the same test}) \\ & - \min(\text{PAC values based on different forms of the same test}).\end{aligned}$$

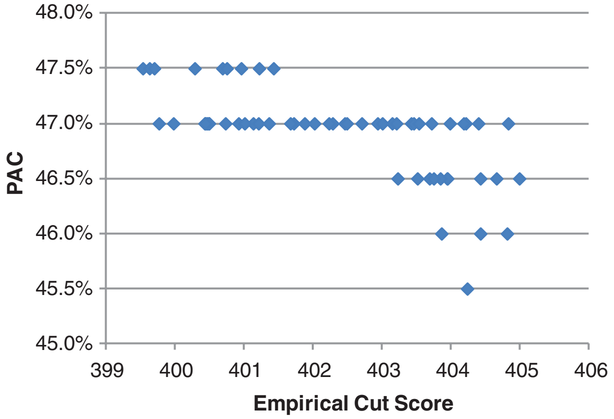

Figure 4 Percentage of examinees at or above the cut scores (PAC) and empirical cut score (test length = 60, theoretical cut score = 400, sample size = 200, *ability* ~ $N(0,1)$, $b_i \sim N(0,1)$).

Table 2 PAC Variation Range

	Theoretical cut score			
Number of items	400	450	500	550
80	1.8	1.7	1.1	0.9
60	2.2	2.1	1.8	1.4
40	3.4	3.0	2.7	2.0
20	6.4	6.2	4.2	3.0

Note. (*ability* ~ $N(0,1)$, $b_i \sim N(0,1)$).

In an ideal situation when the PAC variation range is 0, PAC is not affected by the discreteness in score scale and change in PAC statistics can more accurately indicate progress or regress in performance. Note, however, other nonperformance-related factors may also affect PAC. A nonzero PAC variation range indicates that the PAC statistic may be biased due to the specific test form given. In practice, the PAC values are almost always underestimated because the empirical cut score has to be higher than the theoretical cut score, except for the situation when rounded scale scores are used for reporting.

The PAC variation range is summarized in Table 2 for the first set of simulation studies in which both examinee abilities and item difficulties follow a standard normal distribution. Only the results with examinee sample size of 100,000 are summarized as they are minimally affected by sample variation. The table shows that the PAC variation range is considerably affected by the number of items and moderately affected by the location of the theoretical cut score. When the number of items is larger, the PAC variation range tends to be smaller. This finding is reasonable because the increase in the number of items decreases the discreteness of the reported scale. In theory infinite number of items would reduce the PAC variation range to zero. The pattern observed between the PAC variation range and the theoretical cut score is less straightforward to interpret. The location of the theoretical cut score affects PAC through two other factors: item density for items with difficulties near the specific location and examinee density for examinees with abilities near the specific location. To separately evaluate the effect of item density and examinee density at cut score location on PAC variation range, results from the second and third set of simulations were summarized.

Table 3 presents the PAC variation range values from the second set of simulations in which examinee abilities were generated from a standard normal distribution and item difficulty parameters were generated from a uniform distribution. In this case, it can be assumed that the density of items with difficulties close to the cut score location does not affect the PAC variation range. Table 3 shows a similar pattern with Table 2 in that the PAC variation range decreases with the increase in number of items. With uniform item difficulty distribution, the effect of the theoretical cut score location on PAC is mostly related to examinee density at the cut score point. Therefore the conclusion can be made that the PAC variation range tends to be higher when there is a greater percentage of examinees with abilities near the theoretical cut score. This finding is reasonable as the greater the percentage of examinees near the cut score is, the more examinees there are who can be affected by the difference between the theoretical and empirical cut score. Compared to Table 2,

Table 3 PAC Variation Range

	Theoretical cut score			
Number of items	400	450	500	550
80	2.2	1.9	1.4	1.0
60	2.9	2.4	1.7	1.2
40	4.4	3.6	2.3	1.6
20	9.6	7.0	5.2	3.1

Note. ($ability \sim N(0,1)$, $b_i \sim U(-2,2)$).

Table 4 PAC Variation Range

	Theoretical cut score			
Number of items	400	450	500	550
80	1.1	1.1	1.2	1.7
60	1.4	1.4	1.9	2.6
40	2.2	2.1	2.9	3.5
20	4.2	4.7	4.4	4.9

Note. ($ability \sim U(-2,2)$, $b_i \sim N(0,1)$).

PAC range values are higher in Table 3 under theoretical cut scores of 400 and 450. This difference is because the uniform distribution, compared to standard normal distribution, reduced the number of items that of medium difficulty levels, which subsequently aggravated scale discreteness in that region. For cut scores of 500 and 550 that are further away from the center, the PAC variation range is comparable in Tables 3 and 2. The uniform distribution of the item difficulties does not seem to increase item information sufficiently in the area to reduce discreteness in score scale.

Table 4 summarizes the PAC variation range values under the simulation condition that examinee abilities were generated from a uniform distribution and item difficulty parameters were generated from a standard normal distribution. In this case we can assume that the percentage of examinees with abilities near the cut-score location should not affect the PAC variation range. Consistent with Tables 2 and 3, Table 4 shows that the PAC variation range decreases when the number of items increases. There is also a pattern of decreasing PAC variation range as the theoretical cut score moves from 400 to 550. Given a uniform examinee ability distribution, the effect of the theoretical cut-score location on the PAC variation range is mostly caused by the item difficulty distribution. The results show that PAC variation range tends to be higher when there are fewer items with difficulty levels near the theoretical cut score. The effect of item density at the cut-score level on PAC variation range, however, is less than the effect of examinee density at the cut score. This is demonstrated by relatively small amount of change in the PAC variation range in Table 4 at fixed test length.

Results: PAC in Trend Analyses and Adjustment to PAC Statistics

So far the results have shown how much PAC statistics tend to vary due to discreteness in score scale within a single test administration. In essence, the variation is due to the fact that one cannot administrate test forms that have infinite number of items. Overlooking this effect is less consequential when one is only interested in the PAC statistics within one particular administration. On the other hand, if trend analysis is being conducted using PAC statistics, it is crucial that PAC variation due to discreteness in score scale be considered in the interpretation of results.

For instance, suppose a percentage proficient increase of 3% is observed from Year 1 to Year 2 on a 40-item test, which is not uncommon on K – 12 large-scale state assessments. The plots in Figure 1 on FCAT provide some realistic insight of how much change in PAC on a state assessment is expected from year to year. Suppose that accompanying this percentage proficient increase is a decrease in the empirical cut score from Year 1 to Year 2. And from Table 1 it can be seen that for a 40-item test, the PAC variation due to discreteness in score scale can go well up to 3%. Therefore part or even all of the percentage proficient increase may be due to discreteness in the score scale. Without a procedure such as that outlined in this study, it is difficult to break apart the effect of an increase in performance from the effect of a decrease in the empirical cut score. In another scenario, there may be no change in percentage proficient from Year 1 to Year 2, accompanied by an

Table 5 PAC Trend From Year 1 to Year 5

Theoretical cut	Year	Empirical cut	Observed PAC (%)	Adjusted PAC (%)
400	1	404.16	48.98	50.58
	2	399.78	51.39	51.47
	3	407.54	48.99	52.09
	4	401.71	51.86	52.50
	5	403.21	52.02	53.13
500	1	503.87	15.16	16.06
	2	504.57	16.20	17.34
	3	500.16	17.41	17.45
	4	507.54	16.36	18.12
	5	503.27	17.78	18.66

Note. PAC = percentage of examinees at or above the cut scores.

increase in the empirical cut score. Instead of concluding that there is no performance progression, one should note that performance progression may be disguised by the effect of the empirical cut score increase.

In long-term trend analyses, however, the effect of discreteness in the score scale on PAC statistics is less critical. Because the change in the empirical cut score does not go in a single direction across years, the effect of the difference between the empirical and theoretical cut score does not accumulate across multiple years. The wider the span of test administrations is on which the trend analysis is conducted, the more the effect of discreteness in score scale evens out across administrations and the less the resulting general trend is affected by the issue. The year-to-year (or form-to-form for tests that are administered multiple times in a year) effect can still be observed through zigzags along the general trend.

Given that PAC statistics are often tied to important decisions with regard to educational reform and teaching effectiveness, it cannot be overemphasized that PAC change in trend analyses should be interpreted together with the change in empirical cut scores. Overestimation of performance growth or underestimation of performance regress is expected when PAC change is associated with the decrease of the empirical cut scores. However, underestimation of performance growth or overestimation of performance regress is expected when PAC change is associated with the increase of empirical cut scores.

Aside from being cautious when interpreting change in PAC statistics, it makes sense to adjust PAC statistics so that they are less susceptible to the effect of discreteness in score scales. One convenient way to make the adjustment is to use the interpolated value associated with the theoretical cut score based on the tabled values of PAC statistics against obtainable scale scores. To illustrate the procedure using the hypothetical test in Table 1 as an example, in 2014 the percentage at and above 454 is 54% and the percentage at and above 449 is 58%. Suppose 450 is the theoretical cut score, and therefore the percentage at and above 450 is of interest. Because there is no raw score that corresponds to a scale score of exactly 450, the empirical cut score would be 454 and the unadjusted observed PAC statistic would be 54%. Given that the observed relationship between the empirical cut score and PAC statistics is almost linear within a short span of scores, as observed from Figures 1 to 3, an adjustment can be made by obtaining the interpolated percentage at and above the value that corresponds to a scale score of 450 based on the percentage at and above values associated with scale cores of 449 and 454. This finding leads to an adjusted PAC value of about 57%.

To see how the linear interpolation procedure affects the trend across a number of year, a data set that consists of student responses to a test of 40 items in 5 consecutive years was simulated. The item parameters and examinee abilities follow the same distributions outlined in earlier simulations in this study, except that the mean of the abilities was simulated to be −0.04, −0.02, 0, 0.02, and 0.04 in Years 1 to 5, respectively, to approximate a steady growth in student abilities across years. Two theoretical cuts were used: 400 and 500. Given the simulated ability distributions, PAC is expected to have a yearly increase of about 0.8% with the theoretical cut score being 400 and a yearly increase of about 0.5% with the theoretical cut score being 500. Because test forms differ from year to year, empirical cut scores differ accordingly despite the same theoretical cut scores. Table 5 shows the observed PAC and the adjusted PAC based on interpolation method across the 5 years and the empirical cut scores associated with them. Figures 5 and 6 add the trend line to the observed and adjusted PAC values. The trend line for the adjusted PAC exhibits a smooth growth pattern as expected while the trend for the observed PAC is more erratic.

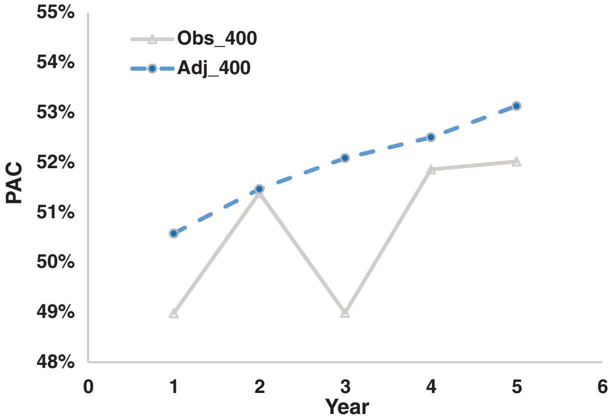

Figure 5 Observed and adjusted percentage of examinees at or above the cut scores (PAC) trend with theoretical cut of 400.

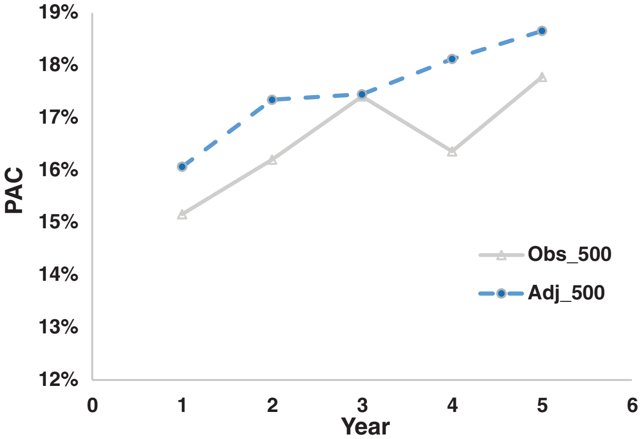

Figure 6 Observed and adjusted percentage of examinees at or above the cut scores (PAC) trend with theoretical cut score of 500.

Alternatively, a more accurate summary of the relationship between the empirical cut score and the percentage above may be achieved by using the jackknife resampling method described in this paper. This summary would give more information on the relationship over the span of all possible empirical cut scores. However, given an almost perfect linear relationship between PAC and the empirical cut score with minimum scatter, adjustment based on linear interpolation, which is much simpler to implement, would probably serve the purpose well.

Discussion

Positive test-score trends are widely interpreted as indicators of educational improvement. In addition to the general interest in the number of schools making adequate yearly progress, there is also immense concern over any unusual performance trend reflected by year-to-year PAC statistics. The change in PAC statistics, particularly the change in the percentage of students at the proficient level, is the most commonly used trend statistic and is associated with important interpretations of increase in student learning, educational improvement, and reform success. Research that examines the variance of PAC statistics due to discreteness in score scales has been scanty. In this paper, results show that discreteness in score scales can affect PAC statistics considerably. The effect is especially significant when the test is short and when there is high examinee density and low item density near the cut score on the reporting score scale. Test length and examinee density at the cut score are found to be more influential than item density at the cut score as factors that tune the effect of discreteness in score scales on PAC statistics.

Although IRT was used to simulate examinee response data and produce scale scores in this study, it should be noted that the method described to evaluate the effect of discreteness in score scale sets no requirement for any specific calibration model or equating method. The fundamental components needed for implementing the jackknife method are observed response data and raw-to-scale score conversion for each jackknife replicate produced using the operational equating method. The linear interpolation method is recommended to adjust PAC statistics when they are used in trend analyses. The method requires only the observed response data and the operational raw-to-scale score conversion. Therefore the methods outlined in this study can be used for any testing program that uses equating to ensure form-to-form comparability, including all classical and IRT equating procedures.

The PAC statistics can be affected by many factors and the discreteness in score scale is just one factor. To attempt to produce precise projections of estimated effects of discreteness on the PAC statistics, some of the factors such as measurement error, calibration error, and equating error were not built into simulations in this study. In reality all these factors may interact with each other and the resulting PAC statistics may be affected by the combination of all these factors. The purpose of this study, however, is to single out the effect of discreteness in score scale. In practice when the suggested methods are used operationally, results will show the variation of the PAC statistics due to discreteness in score scale with the variation caused by other existing factors held constant.

As a result of this study, it is suggested that the change in PAC statistics be interpreted with caution, or that the adjusted PAC statistics be used when the test is high stakes. The adjusted PAC statistics can rule out the effect of discreteness in the score scale so that the change in PAC statistics can more accurately capture progress or regress in student performance. If PAC statistics are not adjusted when evaluating the trend or progress, more detailed information such as how close the empirical cut score is to the theoretical one should be provided. The larger the difference between the empirical cut score and the theoretical one, the more the discreteness in the score scale affects the PAC statistics. Average-based statistics such as changes in scale-score averages and distribution-based statistics such as PP plots provide additional information about the trend and should be viewed together with the PAC statistics in trend analyses.

References

Arce-Ferrer, A., Frisbie, D. A., & Kolen, M. J. (2002). Standard errors of proportions used in reporting changes in school performance with achievement levels. *Educational Assessment*, *8,* 59–75. https://doi.org/10.1207/S15326977EA0801_04

Betebenner, D. W., Shang, Y., Xiang, Y., Zhao, Y., & Yue, X. (2008). The impact of performance level misclassification on the accuracy and precision of percent at performance level measures. *Journal of Educational Measurement, 45,* 119–137. https://doi.org/10.1111/j.1745-3984.2007.00056.x

Florida Department of Education. (n.d.). *FCAT Reading & Mathematics SSS*. Retrieved from http://fcat.fldoe.org/fcinfopg.asp

Haertel, E. H., Thrash, W. A., & Wiley, D. E. (1978). *Metric-free distributional comparisons.* Chicago, IL: ML-Group for Policy Studies in Education.

Ho, A. D. (2007). Discrepancies between score trends from NAEP and state tests: A scale-invariance perspective. *Educational Measurement: Issues and Practice, 26*, 11–20. https://doi.org/10.1111/j.1745-3992.2007.00104.x

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher, 37,* 351–360. https://doi.org/10.3102/0013189X08323842

Ho, A. D. (2009). A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics, 34,* 201–228. https://doi.org/10.3102/1076998609332755

Holland, P. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics, 27*, 3–17. https://doi.org/10.3102/10769986027001003

Kane, T. J., & Staiger, D. O. (2002). Volatility in school test scores: Implications for test based accountability systems. In D. Ravitch (Ed.), *Brookings papers on education policy* (pp. 235–283). Washington, DC: Brookings Institution.

Kolen, M., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer. https://doi.org/10.1007/978-1-4757-4310-4

Koretz, D. (2005). Alignment, high stakes, and the inflation of test scores. *Yearbook of the National Society for the Study of Education, 104*(2), 99–118. https://doi.org/10.1111/j.1744-7984.2005.00027.x

Linn, R. L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis, 24*, 29–36. https://doi.org/10.3102/01623737024001029

Livingston, S. A. (2006). Double P-P plots for comparing differences between two groups. *Journal of Educational and Behavioral Statistics, 31,* 431–435. https://doi.org/10.3102/10769986031004431

Quenouille, M. H. (1949). Problems in plane sampling. *Annals of Mathematical Statistics, 20,* 355–375.. https://doi.org/10.1214/aoms/1177729989

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika, 43,* 353–360. https://doi.org/10.1093/biomet/43.3-4.353

Schafer, W. D., Liu, M., & Wang, H. (2007). Content and grade trends in state assessments and NAEP. *Practical Assessment, Research & Evaluation, 12*(9), 1 – 25.

Spencer, B. D. (1983). On interpreting test scores as social indicators: Statistical considerations. *Journal of Educational Measurement, 20,* 317 – 333. https://doi.org/10.1111/j.1745-3984.1983.tb00210.x

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201 – 210. https://doi.org/10.1177/014662168300700208

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics, 29,* 614 – 623. https://doi.org/10.1214/aoms/1177706647

Wilk, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika, 55,* 1 – 17. https://doi.org/10.1093/biomet/55.1.1

Yen, W. M. (1997). The technical quality of performance assessments: Standard errors of percents of pupils reaching standards. *Educational Measurement: Issues and Practice, 16*(3), 5 – 15. https://doi.org/10.1111/j.1745-3992.1997.tb00594.x

Suggested citation:

Lu, Y. (2017). *Variability in percentage above cut scores due to discreteness in score scales* (Research Report No. RR-17-32). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12158

Action Editor: Gautam Puhan

Reviewers: Tim Davey and Usama Ali