

The Accuracy of Students' Novice, Apprentice, Proficient, and Distinguished Classifications for the 2013 Kentucky End-of-Course Tests

Prepared for :

Kentucky Department of Education
Capital Plaza Tower, 18th Floor
500 Mero Street
Frankfort, KY 40601

Authors: Emily R. Dickinson
Arthur A. Thacker

Date: May 16, 2014



The Accuracy of Students' Novice, Apprentice, Proficient, and Distinguished Classifications for the 2013 Kentucky End-of-Course Tests

Table of Contents

Background.....	1
Methods and Results	1
Discussion and Conclusions	5
References.....	6

List of Tables

Table 1. English 10 Percentages of True Scores Being in Assigned 2013 Classification.....	3
Table 2. Algebra II Percentages of True Scores Being in Assigned 2013 Classification	3
Table 3. Biology Percentages of True Scores Being in Assigned 2013 Classification.....	4
Table 4. US History Percentages of True Scores Being in Assigned 2013 Classification	4
Table 5. Total Percent Expected Correct Classifications and Average Category Distribution Error with Comparison Data from K-PREP 2012.....	5

The Accuracy of Students' Novice, Apprentice, Proficient, and Distinguished Classifications for the 2013 Kentucky End-of-Course Tests

Background

Following adoption of the Kentucky Core Academic Standards (KCAS) through Senate Bill 1 in 2009, an end-of-course (EOC) assessment program at the high school level was approved by the Kentucky Board of Education. ACT's QualityCore® EOC Assessments for English 10, Algebra, II, Biology, and US History (modified version) are administered throughout the school year as students complete the requirements to earn course credit.

ACT provides student-level scores for the multiple-choice portion of the assessments and provides student-level scale scores and the percentile at which their scores rank compared to the national sample. In 2013, the Kentucky Department of Education (KDE) asked the Human Resources Research Organization (HumRRO) to conduct a policy capture focus group among education stakeholders to recommend cut scores for the EOC exams. During this focus group, expert panelists set cut points to allow for the categorization of student scores into performance levels- Novice, Apprentice, Proficient, and Distinguished (NAPD). These cut points were selected to indicate that students scoring at the Apprentice level had a reasonable opportunity, perhaps with supports and/or remediation, of college or career success, that students scoring at the Proficient level were ready for credit bearing college classes, and that students scoring at the Distinguished level had some likelihood of qualifying for academic scholarships (Thacker, Dickinson, & Sinclair, 2013).

Tests are useful when classification accuracy to NAPD levels is high—however, no test is perfect. This report examines the accuracy of the EOC NAPD assignments for 2013.

Methods and Results

Estimating Classification Accuracy

Classical test theory is based on the concept of a “true” score for each examinee, defined as the expected or average score across an infinite number of repeated tests (see, for example, Lord and Novick, 1968). In most cases, we have only a score from a single administration of the test. The difference between this single “observed” score and the underlying “true” score is error. In the present case, we are concerned with the impact of these errors on classifying students into NAPD categories. Specifically, our concern is estimating the rate of classification errors, defined as the probability that students could be classified in an NAPD categories based on a single test other than the one in which their “true” score falls.

Observed scores are assumed to vary around theoretical true scores. This variation of observed scores around true scores is calculated as the standard error of measurement (σ_e). In traditional reliability and generalizability theory, σ_e is a simple function of reliability (r_{tt}) and total test variability (σ_T):

$$\sigma_e = \sigma_T (1-r_{tt}) \quad (1)$$

Error bands around estimated scores often accompany reports of students' test scores. These error bands are based on σ_e with the assumption that errors of measurement are normally

distributed. Although σ_e reflects the variation of observed scores around true scores, because σ_e is constant across the scale of measurement, σ_e is also used to estimate the likelihood of true scores being within predictable intervals around observed scores. For students with a given observed score, the distribution of their true scores can be estimated by σ_e . For example, error bands may be constructed to show the interval around observed scores, which with 95 percent confidence, should contain students' true scores.

If score intervals are divided into proficiency levels, such as NAPD, then σ_e may also be used to estimate misclassification. That is, for any given estimated score, the probability that a student's true score lies in the score interval of an adjacent proficiency classification can be estimated from σ_e and normal distribution cumulative probabilities. For the present study, scale score standard errors of measurement were used to calculate the probability of a student's true score being in the range of scores associated with each of the NAPD categories, given their observed score. This resulted in a distribution of probabilities for each possible scale score and each proficiency category. These probabilities were then weighted by the frequency of students actually scoring at each scale score point. Finally, these weighted values were summed to yield the number of students whose true score would be expected to fall within each proficiency category.

Standard errors of measurement have traditionally been recognized as varying across the distribution of true and observed scores (Brennan, 1998). Scale scores for the QualityCore® were developed using a method that produces standard errors of measurement that are nearly equivalent across the score distribution (ACT, 2010), an approach that enhances interpretability of test scores by establishing a common confidence interval around all student scores (Kolen, 1988). For the present study, classification accuracy was estimated using the scale score standard error of measurement of 2.1 reported in the QualityCore® technical manual (ACT, 2010).

Classification Accuracy Tables

Tables 1 through 4 present classification accuracy estimates for the four EOC assessments used for accountability purposes in Kentucky: English 10, Algebra II, Biology, and US History. In each table, the bold italicized numbers indicate proportions of accurate classifications for each of the NAPD classifications. Using English 10 as an example, 35.61% of students are expected to be accurately classified as Novice, 2.03% are expected to be accurately classified as Apprentice, 38.42% are expected to be accurately classified as Proficient, and 9.12% are expected to be accurately classified as Distinguished. The sum of these four percentages (85.18), labeled "Total % Expected Correct Assignments," is the percent of all students expected to be classified accurately. That is, approximately 85% of all students taking the English 10 EOC assessment would be assigned to the same category of proficiency if we actually knew their true achievement.

The numbers in non-bold italics indicate the proportions of students expected to have true achievement classifications that are different from their assigned classification. For example, 1.45% of all students taking the English 10 assessment are expected to have obtained test scores that place them in the Novice range while their true achievement would place them one category higher in the Apprentice category. Conversely, 2.65% of students are expected to have obtained test scores that place them in the Apprentice category, while their true achievement would place them one category lower in the Novice category. Another 2.75% of students are

expected to have obtained test scores that also place them in the Apprentice category, while their true achievement would place them one category higher in the Proficient category.

English 10 presents a unique situation in which there are higher percentages of misclassification than correct classification for the Apprentice category. EOC cut scores were set by convening an expert panel to identify NAPD cut scores on the ACT reporting scale, which were then linked to the EOC reporting scale through an equipercntile equating process (Thacker, Dickinson, & Sinclair, 2013). This process yielded relatively narrow EOC score ranges for the Apprentice category in particular. In the case of the English 10 EOC, students must score one of only two possible scale scores (152-153) to be classified as Apprentice. Because the standard error of measurement of 2.1 scale score points is slightly larger than the score range for this category, it is reasonable that many students' true scores would fall into an adjacent category.

Table 1. English 10 Percentages of True Scores Being in Assigned 2013 Classification

True Classification	Assigned Classification				Total % Expected
	Novice	Apprentice	Proficient	Distinguished	
Novice	35.61	2.65	0.75	0.00	39.01
Apprentice	1.45	2.03	2.10	0.00	5.58
Proficient	0.56	2.75	38.42	1.63	43.36
Distinguished	0.00	0.00	2.93	9.12	12.05
Total % Assigned	37.62	7.43	44.20	10.75	100
Total % Expected Correct Assignments	85.18% Average Distribution Error:				1.35

Note. Classification accuracy was estimated using the final EOC score data and so distributions of assigned classifications differ slightly from those reported in the preliminary application of EOC cut scores to 2013 data reported by Thacker and Dickinson (2013).

Table 2. Algebra II Percentages of True Scores Being in Assigned 2013 Classification

True Classification	Assigned Classification				Total % Expected
	Novice	Apprentice	Proficient	Distinguished	
Novice	24.49	4.95	0.02	0.00	29.46
Apprentice	3.76	26.32	4.30	0.01	34.39
Proficient	0.02	5.59	18.68	1.40	25.69
Distinguished	0.00	0.02	3.03	7.42	10.47
Total % Assigned	28.27	36.88	26.03	8.83	100
Total % Expected Correct Assignments	76.91% Average Distribution Error:				1.42

Note. Classification accuracy was estimated using the final EOC score data and so distributions of assigned classifications differ slightly from those reported in the preliminary application of EOC cut scores to 2013 data reported by Thacker and Dickinson (2013).

Table 3. Biology Percentages of True Scores Being in Assigned 2013 Classification

True Classification	Assigned Classification				Total % Expected
	Novice	Apprentice	Proficient	Distinguished	
Novice	19.41	3.22	0.00	0.00	22.63
Apprentice	2.90	36.45	3.36	0.00	42.71
Proficient	0.00	3.20	20.77	1.32	25.29
Distinguished	0.00	0.00	2.69	6.67	9.36
Total % Assigned	22.31	42.87	26.82	7.99	100
Total % Expected Correct Assignments		83.30%	Average Distribution Error:		0.85

Note. Classification accuracy was estimated using the final EOC score data and so distributions of assigned classifications differ slightly from those reported in the preliminary application of EOC cut scores to 2013 data reported by Thacker and Dickinson (2013).

Table 4. US History Percentages of True Scores Being in Assigned 2013 Classification

True Classification	Assigned Classification				Total % Expected
	Novice	Apprentice	Proficient	Distinguished	
Novice	30.03	3.62	0.26	0.00	33.91
Apprentice	4.00	6.58	3.27	0.00	13.85
Proficient	0.35	3.65	29.04	2.77	35.81
Distinguished	0.00	0.00	2.69	13.74	16.43
Total % Assigned	34.38	13.85	35.26	16.51	100
Total % Expected Correct Assignments		79.39%	Average Distribution Error:		0.28

Note. Classification accuracy was estimated using the final EOC score data and so distributions of assigned classifications differ slightly from those reported in the preliminary application of EOC cut scores to 2013 data reported by Thacker and Dickinson (2013).

Student classification accuracy data also has important implications for school accountability scores. School accountability scores under the new Unbridled Learning accountability model include achievement and gap components, both of which are a function of student-level classifications. Some of the inevitable classification error will be in one direction and some in the other. As seen in Table 1, some proportion of students are expected be classified higher than their true proficiency and some lower. The percentage of students actually assigned to a particular category versus our projections of the percent of students expected to have true achievement at that level shows that the misclassification errors tend to balance out for the total student population. For example, the last column in Table 1 shows that 5.58% of the students assessed in English 10 are expected to be Apprentice based on their unknowable true scores, while 7.43% of the students are assigned the Apprentice classification based on their test performance. The difference between these percentages is approximately 2. Differences between the expected and assignment distributions for the other three categories are similarly close, with the average difference in category percentages being 1.35. Because this error refers to category total distributions, it is referred to as “Average Distribution Error” in the tables. This can be interpreted as the average difference between expected and observed classifications. These values for each of the EOC assessments are no larger than 1.4, indicating small differences between expected and observed classifications across the NAPD categories.

Discussion and Conclusions

HumRRO routinely calculates estimates of student classification accuracy for the Kentucky state accountability assessments in grades 3 through 8. This is the first such study for the high school EOC assessments. Because the EOC assessments are unique in that they have been scaled so that a single standard error of measurement is used across the scale and across subject matter tests, an additional step was taken to document the appropriateness of classification accuracy estimates for the EOC assessments.

Table 5 compares the percentage of expected correct classifications and average category distribution error for the 2013 EOC assessments with estimates from the 2012 K-PREP assessments in similar content areas. Expected percentages of correct classification are higher for the English 10 and Biology EOC assessments than for their K-PREP counterparts, while Algebra II and US History estimates are slightly lower. Though there is no objective standard identifying acceptable classification accuracy, other state assessments have been found to demonstrate accuracy rates ranging from 61% to 96% (Dickinson, Levinson, Thacker, & Hoffman (2013). Though there is also no objective standard for making judgments about average category distribution errors, those calculated for the EOC assessments tend to be very similar or less than those from the K-PREP assessments. Taken together, these data indicate that of student classification accuracy for the EOC assessments are appropriate.

Table 5. Total Percent Expected Correct Classifications and Average Category Distribution Error with Comparison Data from K-PREP 2012

Total % Expected Correct				Average Category Distribution Error			
KPREP		EOC		KPREP		EOC	
MA 03	78.9	Algebra II	76.9	MA 03	2.1	Algebra II	1.4
MA 04	78.3			MA 04	1.7		
MA 05	79.9			MA 05	1.1		
MA 06	80.1			MA 06	1.5		
MA 07	80.2			MA 07	2.5		
MA 08	80.6			MA 08	1.4		
RD 03	73.4	English 10	85.2	RD 03	0.4	English 10	1.4
RD 04	74.3			RD 04	1.5		
RD 05	74.8			RD 05	1.0		
RD 06	76.9			RD 06	1.4		
RD 07	74.5			RD 07	1.5		
RD 08	75.5			RD 08	1.4		
SC 04	78.1	Biology	83.3	SC 04	2.1	Biology	0.9
SC 07	78.5			SC 07	1.4		
SS 05	80.5	US History	79.4	SS 05	1.7	US History	0.3
SS 08	80.2			SS 08	2.3		

References

- ACT. (2010). *QualityCore® Technical Manual*. Retrieved from <http://www.act.org/qualitycore/pdf/TechnicalManual.pdf>
- Brennan, R. L. (1998). Raw-score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement, 22*, 301-331.
- Dickinson, E. R., Levinson, H., Thacker, A. A., & Hoffman, R. G. (2013). *The Accuracy of Students' Novice, Apprentice, Proficient, and Distinguished Classifications for the 2012 Kentucky Performance Rating for Educational Progress (K-PREP) Tests (2013 No. 037)*. Alexandria, VA: Human Resources Research Organization.
- Kolen, M. J. (1988). Defining scale scores in relation to measurement error. *Journal of Educational Measurement, 25*, 97-110.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Thacker, A. A., & Dickinson, E. R. (2013). *Application of End-of-Course Cut Scores to 2013 End-of-Course (EOC) Student Data (2013 No. 055)*. Alexandria, VA: Human Resources Research Organization.
- Thacker, A. A., Dickinson, E. R., & Sinclair, A. L. (2013). *Policy Capture for Setting End-of-Course and Kentucky Performance Rating for Educational Progress (K-PREP) Cut Scores (2013 No. 007)*. Alexandria, VA: Human Resources Research Organization.