2013 No. 037

The Accuracy of Students' Novice, Apprentice, Proficient, and Distinguished Classifications for the 2012 Kentucky Performance Rating for Educational Progress (K-PREP) 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
R. Gene Hoffman
Heather Levinson

Date: June, 2013

HumRRO®
Human Resources Research Organization

Human Resources Research Organization (HumRRO)
66 Canal Center Plaza, Suite 700, Alexandria, Virginia 22314-1578
Phone: 703.549.3611 | Fax: 703.549.9661 | www.humrro.org

HumRRO
Human Resources Research Organization

2013 No. 037

The Accuracy of Students' Novice, Apprentice, Proficient, and Distinguished Classifications for the 2012 Kentucky Performance Rating for Educational Progress (K-PREP) 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
R. Gene Hoffman
Heather Levinson

Date: June, 2013

THE ACCURACY OF STUDENTS' NOVICE, APPRENTICE, PROFICIENT, AND DISTINGUISHED CLASSIFICATIONS FOR THE 2012 KENTUCTY PERFORMANCE RATING FOR EDUCATIONAL PROGRESS (K-PREP) TESTS

Table of Contents

Classification Accuracy 2
Reading Classification Accuracy Tables: Spring 2012 K-PREP Grade 4 Reading as an Example 2
Summary of the Results 4
References 7

Appendix A Classification Accuracy Tables for K-PREP Spring 2012 1
Appendix B Technical Details 7

List of Tables

Table 1. Grade/Subject Combinations for the K-PREP Test Analyzed for Classification Accuracy 1
Table 2. Grade 4 Reading 2012 Percentages of True Scores Being in Assigned Classification 3
Table 3. Total Percent Expected Correct Classifications and Average Category Distribution Error with Comparison Data from KCCT 2008 4
Table 4. Comparison of K-PREP and Tests Total Percent Expected Correct Classifications With Three Other States 6
Table A-1. Grade 3 Mathematics 2012: Percentages of True Scores Being in Assigned Classification 1
Table A-2. Grade 4 Mathematics 2012: Percentages of True Scores Being in Assigned Classification 1
Table A-3. Grade 5 Mathematics 2012: Percentages of True Scores Being in Assigned Classification 1
Table A-4. Grade 6 Mathematics 2012: Percentages of True Scores Being in Assigned Classification 2
Table A-5. Grade 7 Mathematics 2012: Percentages of True Scores Being in Assigned Classification 2
Table A-6. Grade 8 Mathematics 2012: Percentages of True Scores Being in Assigned Classification 2
Table B-1. Grade 3 Reading 2012: Percentages of True Scores Being in Assigned Classification 3
Table B-2. Grade 4 Reading 2012: Percentages of True Scores Being in Assigned Classification 3
Table B-3. Grade 5 Reading 2012: Percentages of True Scores Being in Assigned Classification 3
Table B-4. Grade 6 Reading 2012: Percentages of True Scores Being in Assigned Classification 4

Table B-5. Grade 7 Reading 2012: Percentages of True Scores Being in Assigned Classification....4
Table B-6. Grade 8 Reading 2012: Percentages of True Scores Being in Assigned Classification....4
Table C-1. Grade 4 Science 2012: Percentages of True Scores Being in Assigned Classification....5
Table C-2. Grade 7 Science 2012: Percentages of True Scores Being in Assigned Classification....5
Table D-1. Grade 5 Social Studies: Percentages of True Scores Being in Assigned Classification....6
Table D-2. Grade 8 Social Studies 2012: Percentages of True Scores Being in Assigned Classification....6

The Accuracy of Students' Novice, Apprentice, Proficient, and Distinguished Classifications for the 2012 Kentucky Performance Rating for Educational Progress (K-PREP) Tests

Background

Following adoption of the Kentucky Core Academic Standards (KCAS) through Senate Bill 1 in 2009, new assessments aligned with KCAS (the Kentucky Performance Rating for Educational Progress, or K-PREP), were developed and administered for the first time in 2012. For scoring and reporting, each grade/subject combination is treated as a separate test. Based on the results of these tests, each student is classified into one of four proficiency levels: Novice, Apprentice, Proficient, or Distinguished (NAPD). The purpose of this report is to present classification accuracy statistics for the Spring 2012 administration of K-PREP. Classification accuracy is a method for considering the reliability of a test.

Grade/subject combinations included in the classification accuracy calculations are presented in Table 1 below. Scoring is a two-step process, as described in the Kentucky Department of Education (KDE) technical manual (KDE, 2012). Students first receive a scale score derived from their responses to the items on the test. "Cut points" have been set to allow for the categorization of student scores into performance levels- Novice, Apprentice, Proficient, and Distinguished (NAPD). New performance level descriptors and cut score standards were developed to indicate mastery levels that are needed to be considered "on track" for college and career readiness.

Table 1. Grade/Subject Combinations for the K-PREP Test Analyzed for Classification Accuracy

Subject[1]	Grades					
	3	4	5	6	7	8
Reading	X	X	X	X	X	X
Mathematics	X	X	X	X	X	X
Science		X			X	
Social Studies			X			X

Tests are useful when classification accuracy to NAPD levels is high—however, no test is perfect. This report examines the accuracy of the K-PREP NAPD assignments for Spring 2012. Results from prior years have been previously reported by Hoffman and Wise (2000b), Hoffman, Wise, and Thacker (2000), Hoffman (2002 and 2003), Hoffman and Dickinson (2004), and Hoffman and Nemeth (2008) and are available at www.kentuckyschools.net (search for "student classification accuracy").

[1] Kentucky also assesses writing in Grades 5, 6, 8, 10, and 11. The writing test is scored using the raw point scale and consists of two writing prompts. No classification statistics were computed for writing. See Thacker, Dickinson & Levinson, 2013 for more information regarding writing scoring accuracy.

Methods and Results

Classification Accuracy

The methodology for this classification accuracy analysis was developed by Hoffman and Wise (1999) and presented to Kentucky's National Technical Advisory Panel on Assessment and Accountability (NTAPAA) on two occasions (September 9-10, 1999 and December 16-17, 1999). The method was approved by the NTAPAA during the September meeting. Preliminary results for the 1999 assessments were presented during the December meeting. The classification accuracy method was also presented to the National Council for Measurement in Education (NCME) at its annual meeting in April 2000. The NCME paper (Hoffman & Wise, 2000a) is available from the authors. While the present report conforms to the NTAPAA reporting specifications and uses the established methodology, we do note that student classification programs were rewritten in 2013 to adjust for changes in the item response model used by the new contractor (Generalized Partial Credit model to a Rasch model) and the removal of multiple-choice and open-response weighting procedures. More detailed technical information is provided in Appendix B.

As mentioned above, no test is perfect, and the reliability of an assessment with its first administration is of particular interest. Reliability of an observed test score is the product of two factors: true proficiency in the knowledge area being assessed and measurement error that comes from a variety of sources. Obtained scores are known, but true scores are unknown. Using test reliability-related statistics, however, it is possible to provide estimations that answer the following two questions:

1. For a given obtained score, what are the odds that true proficiency is in the same NAPD classification?
2. For that given obtained score, what are the odds that true proficiency falls into a different NAPD classification?

These two questions lead to 16 probability estimates: that is, for each of the four assigned NAPD proficiency levels, what are the odds of true proficiency at each of the four NAPD levels? These probability estimates are presented in classification accuracy tables for each grade/subject combination. Classification accuracy tables for the K-PREP are presented in Appendix A.

Reading Classification Accuracy Tables: Spring 2012 K-PREP Grade 4 Reading as an Example

In Table 2 and similar tables in the Appendix, numbers represent percentages of all students, so that the sum of all of the italicized percentages is 100 (within rounding). The "Total % Assigned" row near the bottom of the table indicates the percent of students who were assigned to each of the four NAPD classifications. For example, in Table 2, 32.11% of all Grade 4 students who took the K-PREP in Reading received test scores that placed them in the Novice category. Likewise, 29.55% of all students received test scores within the score range for the Apprentice category, 29.35% were Proficient, and 8.98% were Distinguished.

Since test scores do not perfectly reflect true scores, some proportion of students is expected to have true achievement in categories matching their assigned categories, with the remaining students expected to have true achievement that falls in categories other than their assigned

categories. The bold italicized numbers in Table 2 indicate proportions of accurate classifications. That is, 27.24% of all students are expected to be accurately classified as Novice, 20.66% of all students are expected to be accurately classified as Apprentice, 21.50% of all students are expected to be accurately classified as Proficient, and 4.86% of all students are expected to be accurately classified as Distinguished. The sum of these four percentages (74.26), labeled "Total % Expected Correct Assignments," is the percent of all students expected to be classified accurately. That is, approximately 74% of all Grade 4 students would be assigned to the same category of reading proficiency if we actually knew their true achievement.

Table 2. Grade 4 Reading 2012 Percentages of True Scores Being in Assigned Classification

	Assigned Classification				Total %
True Classification	Novice	Apprentice	Proficient	Distinguished	Expected
Novice	***27.24***	*4.41*	*0.05*	*0.00*	*31.70*
Apprentice	*4.84*	***20.66***	*6.19*	*0.03*	*31.72*
Proficient	*0.03*	*4.48*	***21.50***	*4.09*	*30.10*
Distinguished	*0.00*	*0.00*	*1.61*	***4.86***	*6.48*
Total % Assigned	*32.11*	*29.55*	*29.35*	*8.98*	*100*
Total % Expected Correct Assignments		74.26%	Average Distribution Error:		1.46

The numbers in Table 2 in non-bold italics indicate the proportions of students expected to have true achievement classifications that are different from their assigned classification. For example, 4.84% of all students 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, 4.41% of all 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 4.48% of all 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. In total, 25.74% (100-74.26%) of all students are expected to be misclassified in Grade 4 reading in 2012.

Student classification accuracy data also has important implications for school accountability scores. School accountability scores are a function of all students' classifications. Some of the inevitable classification error will be in one direction and some in the other. As seen in Table 2, 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 2 shows that 31.72% of the students are expected to be Apprentice based on their unknowable true scores, while 29.55% of the students are assigned the Apprentice classification based on their test performance. The difference between these percentages is approximately 2 percent. Differences between the expected and assignment distributions for the other three categories are similarly close, with the average difference in category percentages being 1.46. Because this error refers to category total distributions, it is referred to as "Average Distribution Error" in the table. Results tables for all K-PREP grade/subjects, for 2012, are presented in Appendix A.

Summary of the Results

Table 3 shows the total expected correct assignments for all K-PREP grade/subject combinations for 2012, with the results for KCCT 2008 added for comparison. In 2012, student classification accuracy varies from approximately 73.4% (Grade 3 Reading) to approximately 80.6% (for Grades 8 Mathematics). Overall, Reading has the lowest accuracy with results between 73.4%-76.9%. Mathematics, Science, and Social Studies have accuracies in the 78.1% to 80.6% range. The majority of the accuracy rates for 2012 K-PREP are higher than 2008 KCCT, often by nearly a 10% margin. The most noticeable increase in accuracy can be seen in Grade 4 Science. The only decrease in accuracy was in Grade 7 Reading.

Table 3. Total Percent Expected Correct Classifications and Average Category Distribution Error with Comparison Data from KCCT 2008

	Total Percent Expected Correct Classifications		Average Category Distribution Error	
Subject/Grade	KCCT 2008	K-PREP 2012	KCCT 2008	K-PREP 2012
MA 03		78.9		2.1
MA 04		78.3		1.7
MA 05	70.4	79.9	7.4	1.1
MA 06		80.1		1.5
MA 07		80.2		2.5
MA 08	75.1	80.6	6.6	1.4
RD 03		73.4		0.4
RD 04	73.5	74.3	4.2	1.5
RD 05		74.8		1.0
RD 06		76.9		1.4
RD 07	76.5	74.5	4.1	1.5
RD 08		75.5		1.4
SC 04	67.1	78.1	9.8	2.1
SC 07	72.2	78.5	6.6	1.4
SS 05	72.0	80.5	5.4	1.7
SS 08	70.7	80.2	8.0	2.3

Turning from the individual level accuracy data to the distribution accuracy results, the two right-most columns in Table 3 summarize how well the distribution of assigned classifications matches the expected distribution of true classifications by showing the difference between expected and assigned total percentages averaged across the four achievement levels. As would be expected, this average error is larger among the subject areas with lower level accuracy values. This level of error has decreased substantially since 2008.

Discussion and Conclusions

Test specialists are well aware of the need to study classification accuracy as well as more traditional measures of test reliability. Several methodological papers focusing on analytical variations of the accuracy theme have used operational data. For example, Rogosa (1994) examined California's 1993 CLAS assessment, which uses six proficiency levels. He found that although the probability of classification within one category of true proficiency was nearly 95%, the probability of exact classification was only 51.72%.

In another example, Lee, Hanson, and Brennan (2000) used data from ACT's Work Keys assessment. Their results confirm that the number of proficiency categories makes a difference – more categories mean more opportunities for classification error. For a Work Keys subtest with five categories, exact accuracy for several different forms was in the 70% range, while a subtest with six categories showed accuracy in the low- to mid-60% range. Lee et al. also looked at accuracy for classifying students simply above or below a single cutpoint, using each of the possible Work Keys cutpoints to look at these dichotomous classifications. Accuracy was in the upper 80% range to near 100% for classifying students into only one of two categories. The higher levels of accuracy occurred for classification of students into either extreme. When the cutpoint was closer to the center, accuracy tended to be in the upper 80% range. Young and Yoon (1998) provide similar data from the New Standards assessments. Again, when making only a dichotomous (two-category) classification, they showed better accuracy (e.g., in the lower 90% range).

Table 4 on the next page presents the total percent of students that would be expected to be correctly classified in Kentucky, along with percentages for three other states' past test administrations. Whereas Kentucky categorizes students into four achievement levels, California (Educational Testing Service, 2013) and Florida (HumRRO & Harcourt, 2007) use five. Texas (Texas Education Agency, 2010) reports student classification accuracy as the percentage correctly classified as Meeting Standard. As the table shows, Kentucky students are correctly classified at a rate similar to those found in other states. As expected, the state with the highest classification accuracy reported the percentage accurately classified into two possible categories.

Table 4. Comparison of K-PREP and Tests Total Percent Expected Correct Classifications With Three Other States

Grade	Subject	Kentucky	California[1]	Florida[1]	Texas[2]
Grade 3	Reading[3]	73.4	76.0	71.8	80.1
	Math	78.9	81.0	70.1	80.9
Grade 4	Reading	74.3	81.0	66.6	81.6
	Math	78.3	82.0	67.4	81.6
	Science	78.1			
Grade 5	Reading	74.8	81.0	65.9	82.0
	Math	79.9	80.0	66.6	81.1
	Science		77.0		77.0
	Social Studies	80.5			
Grade 6	Reading	76.9	79.0	96.4	80.8
	Math	80.1	79.0	60.8	83.2
Grade 7	Reading	74.5	78.0	70.0	81.5
	Math	80.2	77.0	62.1	85.8
	Science	78.5			
Grade 8	Reading	75.5	79.0	64.2	80.3
	Math	80.6	74.0	61.8	84.7
	Science		77.0		82.1
	Social Studies	80.2	77.0		85.5

[1] Classification accuracy using five categories.
[2] Classification accuracy using two categories (Met Standard/Did Not).
[3] ELA is tested in California.

It is also possible to calculate accuracy for the K-PREP Grade 4 Reading assessment as if it were used to divide students into two categories created by combining Novice with Apprentice and Proficient with Distinguished. Looking at the data in Table 2 from this perspective, some of the cells that previously represented misclassification now represent accurate classification. Accuracy, therefore, becomes the sum of the four cells in the upper left plus the sum of the four cells in the lower right. The resulting dichotomous accuracy of “Below Proficient” versus “Proficient and Above” is approximately 89.2% which is comparable to Work Keys and New Standards. Hoffman, Diaz, and Dickinson presented similar results with data from the Idaho Standards Achievement Tests (2005). When students were classified as ‘Proficient or Not’, average classification accuracy across all tests jumped from 77.8% to 90.8%. This is particularly pertinent given that Kentucky uses a dichotomous classification system with the goal of increasing student performance to Proficient or greater as a part of the Unbridled Learning accountability system for students, schools, and districts.

Given these examples, K-PREP appears to have classification accuracy statistics that are similar to other educational proficiency assessments and slightly more accurate than KCCT results from 2008. We have also seen in this report that individual level inaccuracies tend to cancel out so that the distributions of students’ scores on the aggregate level appear to be reasonably precise.

References

Educational Testing Service. (2013). *California Standards Test Technical Report Spring 2012 Administration*. Retrieved on June 17, 2013 from http://www.cde.ca.gov/ta/tg/sr/documents/cst12techrpt.pdf#search=STARS%20student%20classification%20accuracy&view=FitH&pagemode=none.

Hoffman, R. G., Diaz, T.E., & Dickinson E. (2005). *Idaho Standards Achievement Test: Independent calculations of reliability estimates, standard errors of measurement,classification accuracy, and classification consistency*. (FR-04-87). Alexandria, VA: Humans Resources Research Organization.

Hoffman, R. G. (2003). *The accuracy of students' novice, apprentice, proficient, and distinguished classifications for the 2003 Kentucky Core Content Tests* (FR-03-73). Alexandria, VA: Human Resources Research Organization.

Hoffman, R. G. (2002). *The accuracy of students' novice, apprentice, proficient, and distinguished classifications for the 2001 and 2002 Kentucky Core Content Tests* (FR-02-46). Alexandria, VA: Human Resources Research Organization.

Hoffman, R.G., & Dickinson, E.R. (2004). *The accuracy of students' novice, apprentice, proficient, and distinguished classifications for the 2004 Kentucky Core Content Tests* (FR-04-77). Alexandria, VA: Human Resources Research Organization.

Hoffman, R. G., & Nemeth, Y. M. (2008). *The accuracy of students' novice, apprentice, proficient, and distinguished classifications for the 2008 Kentucky Core Content Tests* (FR-08-108). Alexandria, VA: Human Resources Research Organization.

Hoffman, R. G., & Wise, L. L. (1999). *Establishing the reliability of student level classifications: Analytic plan and demonstration.(*FR-WATSD-99-34). Alexandria, VA: Human Resources Research Organization.

Hoffman, R. G., & Wise, L. L. (2000a). *Establishing the reliability of student proficiency classifications: The accuracy of observed classifications.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, April, 2000. Also available at www.Humrro.org .

Hoffman, R. G., & Wise, L. L. (2000b). *The accuracy of students' novice, apprentice, proficient, and distinguished classification of the Kentucky Core Content Test for 1999.* (FR-WATSD-00-25). Alexandria, VA: Human Resources Research Organization.

Hoffman, R. G., Wise, L. L., & Thacker, A.A. (2000). *The accuracy of students' novice, apprentice, proficient, and distinguished classification of the 2000 Kentucky Core Content Test for 1999.* (FR-00-41). Alexandria, VA: Human Resources Research Organization.

Human Resources Research Organization and Harcourt Educational Measurement (2004). *Florida Comprehensive Assessment Test for reading and mathematics: Technical report for test administrations of FCAT 2004.* San Antonio, TX: Harcourt Educational Measurement.

Kentucky Department of Education (2012). *Kentucky Performance Rating for Educational Progress 2011-12 technical report.* Frankfort, KY: Author.

Lee, W., Hanson, B. A., & Brennan, R. L. (2000). *Procedures for computing classification consistency and accuracy indices with multiple categories.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, April, 2000.

Rogosa, D. (1994). Misclassification in student performance levels. In CTB/McGraw-Hill. (1994). *1994 CLAS Assessment Technical Report*. Monterrey, CA: Author.

Texas Education Agency. (2010). *Technical Digest for the Academic Year 2008-2009*. Retrieved on June 17, 2013 from http://www.tea.state.tx.us/index3.aspx?id=2147484418&menu_id=793.

Thacker, A. A., Dickinson, E. R. & Levinson, H. (2013). *Exploration of the scoring of the K-PREP on-demand writing assessment*. (2013 No. 038). Alexandria, VA: Human Resources Research Organization.

Young, M. J. & Yoon, B. (1998). *Estimating the consistency and accuracy of classifications in a standards-referenced assessment.* (CSE Technical Report 475). Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Appendix A

Classification Accuracy Tables for K-PREP Spring 2012

Table A-1. Grade 3 Mathematics 2012: Percentages of True Scores Being in Assigned Classification

	Assigned Classification				
True Class	Novice	Apprentice	Proficient	Distinguished	Total % Expected
Novice	***23.76***	*3.58*	*0.00*	*0.00*	*27.35*
Apprentice	*4.34*	***30.45***	*6.94*	*0.00*	*41.73*
Proficient	*0.00*	*3.42*	***22.97***	*2.36*	*28.75*
Distinguished	*0.00*	*0.00*	*.47*	***1.70***	*2.17*
Total % Assigned	28.10	37.45	30.39	4.06	100
Total % Expected Correct Assignments: **78.88**			Average Distribution Error: 2.14		

Table A-2. Grade 4 Mathematics 2012: Percentages of True Scores Being in Assigned Classification

	Assigned Classification				
True Class	Novice	Apprentice	Proficient	Distinguished	Total % Expected
Novice	***25.99***	*3.81*	*0.00*	*0.00*	*29.80*
Apprentice	*6.68*	***28.91***	*3.54*	*0.00*	*39.13*
Proficient	*0.01*	*4.89*	***19.26***	*1.69*	*25.85*
Distinguished	*0.00*	*0.00*	*1.09*	***4.13***	*5.22*
Total % Assigned	32.67	37.61	23.89	5.82	100
Total % Expected Correct Assignments: **78.29**			Average Distribution Error: 1.74		

Table A-3. Grade 5 Mathematics 2012: Percentages of True Scores Being in Assigned Classification

	Assigned Classification				
True Class	Novice	Apprentice	Proficient	Distinguished	Total % Expected
Novice	***24.53***	*3.88*	*0.00*	*0.00*	*28.41*
Apprentice	*6.00*	***35.16***	*3.72*	*0.00*	*44.88*
Proficient	*0.00*	*4.23*	***17.49***	*1.17*	*22.89*
Distinguished	*0.00*	*0.00*	*1.10*	***2.72***	*3.82*
Total % Assigned	30.53	43.27	22.31	3.89	100
Total % Expected Correct Assignments: **79.89**			Average Distribution Error: 1.10		

Table A-4. Grade 6 Mathematics 2012: Percentages of True Scores Being in Assigned Classification

	Assigned Classification				
True Class	Novice	Apprentice	Proficient	Distinguished	Total % Expected
Novice	***20.45***	*2.77*	*0.04*	*0.00*	*23.26*
Apprentice	*5.29*	***15.33***	*4.49*	*0.00*	*25.11*
Proficient	*0.09*	*4.52*	***40.96***	*1.55*	*47.11*
Distinguished	*0.00*	*0.00*	*1.15*	***3.37***	*4.52*
Total % Assigned	25.83	22.62	46.63	4.92	100
Total % Expected Correct Assignments: **80.11**			Average Distribution Error: 1.49		

Table A-5. Grade 7 Mathematics 2012: Percentages of True Scores Being in Assigned Classification

	Assigned Classification				
True Class	Novice	Apprentice	Proficient	Distinguished	Total % Expected
Novice	***26.08***	*2.85*	*0.00*	*0.00*	*28.93*
Apprentice	*7.47*	***28.60***	*3.15*	*0.00*	*39.21*
Proficient	*0.00*	*3.59*	***21.27***	*1.50*	*26.36*
Distinguished	*0.00*	*0.00*	*1.22*	***4.27***	*5.50*
Total % Assigned	33.55	35.04	25.64	5.77	100
Total % Expected Correct Assignments: **80.23**			Average Distribution Error: 2.45		

Table A-6. Grade 8 Mathematics 2012: Percentages of True Scores Being in Assigned Classification

	Assigned Classification				
True Class	Novice	Apprentice	Proficient	Distinguished	Total % Expected
Novice	***24.5***	*3.62*	*0.00*	*0.00*	*28.12*
Apprentice	*5.25*	***31.46***	*3.44*	*0.00*	*40.15*
Proficient	*0.00*	*4.26*	***21.49***	*1.97*	*27.73*
Distinguished	*0.00*	*0.00*	*0.84*	***3.16***	*4.00*
Total % Assigned	29.75	39.34	25.78	5.13	100
Total % Expected Correct Assignments: **80.62**			Average Distribution Error: 1.38		

Table B-1. Grade 3 Reading 2012: Percentages of True Scores Being in Assigned Classification

	Assigned Classification				
True Class	Novice	Apprentice	Proficient	Distinguished	Total % Expected
Novice	***27.91***	*5.23*	*0.13*	*0.00*	*33.28*
Apprentice	*4.42*	***17.23***	*5.72*	*0.02*	*27.39*
Proficient	*0.09*	*5.09*	***21.39***	*3.19*	*29.75*
Distinguished	*0.00*	*0.01*	*2.76*	***6.82***	*9.58*
Total % Assigned	32.42	27.56	29.99	10.03	100
Total % Expected Correct Assignments: **73.35**			Average Distribution Error: 0.43		

Table B-2. Grade 4 Reading 2012: Percentages of True Scores Being in Assigned Classification

	Assigned Classification				
True Class	Novice	Apprentice	Proficient	Distinguished	Total % Expected
Novice	***27.24***	*4.41*	*0.05*	*0.00*	*31.70*
Apprentice	*4.84*	***20.66***	*6.19*	*0.03*	*31.72*
Proficient	*0.03*	*4.48*	***21.50***	***4.09***	*30.10*
Distinguished	*0.00*	*0.00*	*1.61*	*4.86*	*6.48*
Total % Assigned	32.11	29.55	29.35	8.98	100
Total % Expected Correct Assignments: **74.26**			Average Distribution Error: 1.46		

Table B-3. Grade 5 Reading 2012: Percentages of True Scores Being in Assigned Classification

	Assigned Classification				
True Class	Novice	Apprentice	Proficient	Distinguished	Total % Expected
Novice	**32.1**	*4.99*	*0.14*	*0.00*	37.23
Apprentice	*3.91*	**15.49**	*5.92*	*0.03*	25.35
Proficient	*0.07*	*4.41*	**21.98**	*3.86*	30.31
Distinguished	*0.00*	*0.00*	*1.89*	**5.22**	7.11
Total % Assigned	36.08	24.89	29.93	9.10	100
Total % Expected Correct Assignments: **74.79**			Average Distribution Error: 1.00		

Table B-4. Grade 6 Reading 2012: Percentages of True Scores Being in Assigned Classification

	Assigned Classification				
True Class	Novice	Apprentice	Proficient	Distinguished	Total % Expected
Novice	***33.45***	*4.70*	*0.10*	*0.00*	*38.25*
Apprentice	*3.76*	***15.6***	*6.09*	*0.01*	*25.46*
Proficient	*0.04*	*3.36*	***23.54***	*2.84*	*29.77*
Distinguished	*0.00*	*0.00*	*2.18*	***4.34***	*6.52*
Total % Assigned	37.24	23.66	31.91	7.18	100
Total % Expected Correct Assignments: **76.92**			Average Distribution Error: 1.40		

Table B-5. Grade 7 Reading 2012: Percentages of True Scores Being in Assigned Classification

	Assigned Classification				
True Class	Novice	Apprentice	Proficient	Distinguished	Total % Expected
Novice	***26.05***	*4.58*	*0.09*	*0.00*	*30.72*
Apprentice	*3.94*	***17.08***	*5.78*	*0.02*	*26.82*
Proficient	*0.05*	*4.36*	***22.62***	*4.85*	*31.88*
Distinguished	*0.00*	*0.00*	*1.82*	***8.76***	*10.58*
Total % Assigned	30.04	26.03	30.31	13.63	100
Total % Expected Correct Assignments: **74.51**			Average Distribution Error: 1.52		

Table B-6. Grade 8 Reading 2012: Percentages of True Scores Being in Assigned Classification

	Assigned Classification				
True Class	Novice	Apprentice	Proficient	Distinguished	Total % Expected
Novice	***29.76***	*3.86*	*0.06*	*0.00*	*33.67*
Apprentice	*4.91*	***16.93***	*4.89*	*0.01*	*26.75*
Proficient	*0.07*	*4.85*	***21.53***	*3.79*	*30.24*
Distinguished	*0.00*	*0.00*	*2.09*	***7.24***	*9.34*
Total % Assigned	34.74	25.65	28.56	11.05	100
Total % Expected Correct Assignments: **75.46**			Average Distribution Error: 1.39		

Table C-1. Grade 4 Science 2012: Percentages of True Scores Being in Assigned Classification

	Assigned Classification				
True Class	Novice	Apprentice	Proficient	Distinguished	Total % Expected
Novice	***6.55***	*1.77*	*0.00*	*0.00*	*8.32*
Apprentice	*2.23*	***22.12***	*4.85*	*0.00*	*29.21*
Proficient	*0.00*	*4.08*	***34.91***	*6.29*	*45.28*
Distinguished	*0.00*	*0.00*	*2.64*	***14.55***	*17.19*
Total % Assigned	8.78	27.97	42.41	20.84	100
Total % Expected Correct Assignments: **78.13**			Average Distribution Error: 2.06		

Table C-2. Grade 7 Science 2012: Percentages of True Scores Being in Assigned Classification

	Assigned Classification				
True Class	Novice	Apprentice	Proficient	Distinguished	Total % Expected
Novice	***11.65***	*2.81*	*0.00*	*0.00*	*14.46*
Apprentice	*3.33*	***21.74***	*4.50*	*0.00*	*29.57*
Proficient	*0.01*	*5.05*	***36.87***	*4.01*	*45.93*
Distinguished	*0.00*	*0.00*	*1.78*	***8.26***	*10.04*
Total % Assigned	14.98	29.59	43.16	12.27	100
Total % Expected Correct Assignments: **78.50**			Average Distribution Error: 1.39		

Table D-1. Grade 5 Social Studies: Percentages of True Scores Being in Assigned Classification

	Assigned Classification				
True Class	Novice	Apprentice	Proficient	Distinguished	Total % Expected
Novice	***12.44***	*2.43*	*0.00*	*0.00*	*14.87*
Apprentice	*3.20*	***26.48***	*4.61*	*0.00*	*34.29*
Proficient	*0.00*	*4.10*	***36.23***	*3.88*	*44.20*
Distinguished	*0.00*	*0.00*	*1.31*	***5.33***	*6.65*
Total % Assigned	15.64	33.01	42.15	9.21	100
Total % Expected Correct Assignments: **80.48**			Average Distribution Error: 1.67		

Table D-2. Grade 8 Social Studies 2012: Percentages of True Scores Being in Assigned Classification

	Assigned Classification				
True Class	Novice	Apprentice	Proficient	Distinguished	Total % Expected
Novice	***8.89***	*2.08*	*0.00*	*0.00*	*10.97*
Apprentice	*3.08*	***25.66***	*3.88*	*0.00*	*32.62*
Proficient	*0.00*	*3.75*	***34.91***	*5.27*	*43.94*
Distinguished	*0.00*	*0.00*	*1.73*	***10.74***	*12.47*
Total % Assigned	11.97	31.49	40.53	16.01	100
Total % Expected Correct Assignments: **80.20**			Average Distribution Error: 2.27		

Appendix B

Technical Details

Although student scoring for the Kentucky Core Content for Assessment will rest on IRT scaling, the end result for each student will be a classification of their test performance into one of four proficiency levels: Novice, Apprentice, Proficient, or Distinguished (NAPD). Furthermore, school accountability indexes are calculated from these NAPD classifications rather than from student IRT-based scale scores. Therefore, analysis of the potential for student classification errors is important at both the student and school levels of analysis.

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 have a score from a single administration of the test in question. The difference between this single "observed" score and the underlying "true" score is error. In the present case, we are concerned not just with the size of these errors, but with the impact of these errors on classifying students into NAPD categories. Our concern is estimating the rate of classification errors, defined as the probability that students could be classified in an NADP categories based on a single test other than the one in which their true score falls. Classification *accuracy* is usually used to connote the likelihood that classification from observed scores agrees with classification expected by true scores. The converse, which we will call *misclassification*, is the likelihood that a student's true score is in a proficiency classification different from his/her observed score. The phrasing of this last sentence is important because it expresses errors in a meaningful way for individual students and their parents. Parents know their students' test scores; information should be provided which helps them with the uncertainty about their students' true proficiency.

Test theory is concerned with the uncertainty of estimating students' true achievement levels based on their fallible, observed test scores. Observed scores are assumed to vary in lawful ways around theoretical true scores or around domain scores (theoretical scores derived from all possible forms and conditions for measuring performance). This variation of observed scores around true or domain 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):

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

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 probably 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. An obvious implication of this is that the closer any student's observed score is to a cut point between two proficiency levels, the greater the chances that the student is misclassified. Indeed, the likelihood of misclassification approaches 50 percent for scores very near a cut point, regardless of σ_e.

IRT is one of several more elaborate methods that produce estimates of standard errors of measurement that vary along the true ability scale (Feldt and Brennan, 1989). That is, standard errors of measurement are conditioned on student ability. While estimates of conditional standard errors of measurement, i.e., $\sigma(x|\theta)$, increase precision in understanding the relationship between estimated scores and true scores, they create a complication for estimating classification accuracy. Misclassification is based on the distribution of true scores around observed scores, i.e., $\sigma(\theta|x)$, whereas $\sigma(x|\theta)$ represents the opposite – the distribution of observed scores around true scores. When $\sigma(x|\theta)$ varies across ability levels, it cannot confidently be used as a substitute for $\sigma(\theta|x)$.

For each content area, HumRRO produced an estimate of the expected raw score for each true ability (theta). Next, HumRRO calculated the probability of correct item responses (or the probability of each possible score point for constructed response) for each item at each possible theta. For multiple choice items, the formula applied was the following:

$$p1= e^{th(i)-b_hum(j)}/1+e^{th(i)-b_hum(j)}, \qquad (2)$$

where p1 is the probability of a correct item response, th(i) is the ability of person I and b_hum is the difficulty of item j. For short answer items, the formula applied was the following:

$$p0=e^{th(i)-D(0)}/e^{th(i)-D(0)} + e^{th(i)-D(0) + th(i)-D(1)} + e^{th(i)-D(0) + th(i)-D(1) + th(i)-D(2)} \qquad (3)$$

$$p1=e^{th(i)-D(0) + th(i)-D(1)}/e^{th(i)-D(0)} + e^{th(i)-D(0) + th(i)-D(1)} + e^{th(i)-D(0) + th(i)-D(1) + th(i)-D(2)}$$

$$p2=e^{th(i)-D(0) + th(i)-D(1) + th(i)-D(2)}/e^{th(i)-D(0)} + e^{th(i)-D(0) + th(i)-D(1)} + e^{th(i)-D(0) + th(i)-D(1) + th(i)-D(2)},$$

where p0, p1, and p2 are the probabilities of each possible score point, respectively, and D(0), D(1), and D(2) represent the step difficulty of each possible score point, which are essentially points on the x-axis at which the probability curves for each score point intersect. These step difficulty values are expected to increase for each subsequent score point. Similar equations were used for extended response options, but were extended to include 3 and 4 as possible score points.

Based on the item probabilities generated, probabilities for total test raw scores were next generated for each level of theta. Next, probabilities were generated for each possible theta on the 100-300 reporting scale. These two sets of probabilities were combined to produce a joint probability, or $P(ObS|\theta)$. By applying Bayes' Theorem, a probability density for true scale scores for a given observed score, $f(\theta|ObS)$,can be estimated from $P(ObS|\theta)$ and some assumptions about the overall distribution of θ.

Bayes' Theorem, as applied to continuous variables, states that:

$$f(\theta|ObSj) = \frac{f(\theta)P(ObSj|\theta)}{P(ObSj)} \tag{4}$$

where $P(ObSj) = \int P(ObSj|\theta)\, f(\theta)\, d\theta$ and $f(\theta)$ is the density function for the distribution of true scale scores.

To simplify both the computation and the presentation of the proposed approach, true ability is converted to a discrete variable with levels corresponding to the possible scale score estimates. Thus, $P(\theta_\iota) = P\{(ObS_{i-1} + ObS_i)/2 \leq \theta < (ObS_i + ObS_{i+1})/2\}$ where $-\infty$ and ∞ are substituted for ObS_0 and ObS_{k+1} respectively. With this change, the probability of different true scores, θ_i for any given observed score, ObS_j, can be rewritten as

$$P(\theta_i|ObS_j) = \frac{P(ObS_j|\theta_i)P(\theta_i)}{P(ObS_j|\theta_1)P(\theta_1)+P(ObS_j|\theta_2)P(\theta_2)+P(ObS_j|\theta_3)P(\theta_3)+\ldots+P(ObS_j|\theta_k)P(\theta_k)} \tag{5}$$

Where ObS_j = observed scale score at level *j*, with *k* possible values in the scoring table, and

θ_i = true ability at level i, with *k* levels of ability represented in the scoring table.

For each θ_i and its associated $\sigma(x|\theta_i)$, the probability of obtaining a given ObS_j, denoted $P(ObS_j|\theta_i)$, can be calculated directly from the IRT model and item parameter estimates or estimated using $\sigma(x|\theta_i)$ and assuming a normal distribution of errors.

For this report, we used the assumption of normally distributed errors. $P(ObS_j|\theta_i)$ was estimated for each of the *k* times *k* combinations of possible observed scores and true scores by assuming that each discrete scale score includes a hypothetical range of scale values from half of the distance to the next lower possible value to half the distance to the next higher value, i.e.,

$$\text{Score Range for } ObS_j = \frac{ObS_j + ObS_{(j-1)}}{2} \text{ to } \frac{ObS_j + ObS_{(j+1)}}{2} \tag{6}$$

For each θ_i and its associated $\sigma(x|\theta_i)$, the cumulative probability of scores within the score range for ObS_j can be calculated to estimate $P(ObS_j|\theta_i)$, assuming a normal distribution of errors. Raw-score-to-scale-score tables also limit the extremes of the distribution. For the lowest score level, $P(ObS_1|\theta_i)$ is calculated as the cumulative probability of $\frac{ObS_1 + ObS_2}{2}$, given θ_i and $\sigma(x|\theta_i)$. For the highest score level, $P(ObS_k|\theta_i)$ is calculated as the 1 - the cumulative probability of $\frac{ObS_{k-1} + ObS_k}{2}$. given θ_i and $\sigma(x|\theta_i)$.

The probability of misclassification is the probability that students with a given observed score, ObS_j, could have a true score in a proficiency (NAPD) level that is different than the level that contains that ObS_j. This probability can be calculated as

P(Proficiency Level for $\theta \neq$ Proficiency Level from ObS_j)

$$= P(\text{Proficiency Level for } \theta < \text{Proficiency Level from } ObS_j)$$

$$+ \; P(\text{Proficiency Level for } \theta > \text{Proficiency Level for } ObS_j)$$

$$= \sum_{i=1}^{m} P(\theta_i|ObS_j) + \sum_{i=n}^{k} P(\theta_i|ObS_j), \qquad (7)$$

where *m* represents the highest level of θ in the next lower proficiency level from the observed score, *n* represents the lowest level of θ in the next higher proficiency level above the observed score, and *k* represents the highest θ represented in the scoring table.

Equation (7) provides the classification error information that will be the most meaningful for students interpreting their scores. If desirable, Equation (7) could be split into separate estimates of the probability that a student's true classification is lower or is higher than indicated by his/her observed scale score, using only the first term or the second term of Equation (7), respectively. The latter probability may be of more interest to students and parents.

The decision to place scores on transcripts, however, will not be made individually, but will be made on a system-wide level affecting all students. A system level estimate of the proportion of all students expected to be misclassified can be calculated by weighting the results of Equation (7) for each score level with the proportion of the sample who receive that score, and then summing overall score levels. That is,

Proportion of all student expected to be misclassified =

$$\sum_{j=1}^{k} \left\{ \left[\sum_{i=1}^{m} P(\theta_i|ObS_j) + \sum_{i=n}^{k} P(\theta_i|ObS_j) \right] * \frac{Freq_j}{\text{Total of All Students}} \right\}, \qquad (8)$$

where *m* and *n*, defined as above, will vary with the proficiency level of Obs_j.