

An Exploration of Alternate Methods for Scoring and Estimating Item Parameters for the Kentucky Writing Assessments

Arthur A. Thacker
Bethany H. Bynum

Prepared for: Kentucky Department of Education
500 Mero Street
Frankfort, KY 40601

February 18, 2010

An Exploration of Alternate Methods for Scoring and Estimating Item Parameters for the Kentucky Writing Assessments

**Arthur A. Thacker
Bethany H. Bynum**

Prepared for: Kentucky Department of Education
500 Mero Street
Frankfort, KY 40601

February 18, 2010

**AN EXPLORATION OF ALTERNATE METHODS FOR SCORING AND ESTIMATING
ITEM PARAMETERS FOR THE KENTUCKY WRITING ASSESSMENTS**

Table of Contents

Introduction.....2

Validity Concerns.....2

Equating Options4

Scoring Options.....6

Discussion and Conclusions1

Recommendations2

References.....4

AN EXPLORATION OF ALTERNATE METHODS FOR SCORING AND ESTIMATING ITEM PARAMETERS FOR THE KENTUCKY WRITING ASSESSMENTS

Introduction

Kentucky assesses writing in multiple ways in Grades 5 and 8. For many years, students have completed a writing portfolio during the course of an academic year, or even across multiple years. The portfolio was designed to represent the best possible example of a student's writing capability. It was often edited by peers, rewritten multiple times by a student, word-processed, and polished over the course of many days of effort.

Students are also assessed using an on-demand assessment. The on-demand writing assessment is given during the regular Kentucky Core Content Test (KCCT) testing session and is written long hand, in a single testing session. Prior to 2007, the on-demand assessment consisted of a single holistically scored writing prompt. Students could choose one of two prompts to write toward. The scoring rubric was designed such that the Novice, Apprentice, Proficient, and Distinguished (NAPD) performance categories were assigned directly by trained scorers. Equating to prior years relied exclusively on the consistency of the rubric for assigning scores.

Kentucky's on-demand writing assessments in Grades 5 and 8 currently contain a section of multiple-choice items and a writing prompt (students still select one of two). These test items are used to create analytic scores for students. Each student receives subscores for four components of writing, all based on the same set of test items. The four components are process, content, structure, and conventions. Process is measured using the multiple-choice items only. The other three components are based on the writing prompt. Each prompt response is scored separately for the three components.

For the past two years, there have been concerns regarding the validity of the writing subscores (scores for process, content, structure, and conventions). These concerns stem from design issues inherent in the way the assessment is structured, scoring concerns, and analyses of student data. This report will briefly describe those concerns and recommend ways of investigating potential solutions. This is a particularly important endeavor due to Senate Bill 1 (SB1) eliminating writing portfolios as part of the statewide assessment system. In the absence of writing portfolios, the on-demand writing assessment is now the only formal assessment of writing. Consequently, this underscores the importance of ensuring that meaningful scores come out of the on-demand writing assessment.

Validity Concerns

First, the writing assessment is designed such that dimensionality and potential method effects are intertwined and difficult to estimate. The multiple-choice items are all used to measure the process component, and that component is different from the three components measured by the writing prompt. On most assessments, we expect that the

multiple-choice items and the open-response items will correlate well with one another. We expect this because the underlying unidimensionality assumptions states that the construct we are measuring (e.g. writing) is not simply a set of unrelated concepts but a single measurable construct that may contain various interrelated skills and knowledge requirements. This assumption would suggest that the four subscores of writing all measure an underlying overall dimension of writing, allows us to generate plausible overall scores for students in writing. Kentucky's current writing test seems to break this basic assumption. The multiple-choice items are not sufficiently related to the prompt scores for the item parameter estimation software to estimate them together as designed. However, it is impossible to say whether this issue is due to dimensionality inherent in the difference between writing process and the other three components, or if the issue stems from a method effect between multiple-choice items and writing prompts, or some combination of the two.

The potential dimensionality issue does not end with the lack of coherence between the multiple-choice items and writing prompts. It is not uncommon, despite unidimensionality assumptions, for assessments to produce subscores for components within a particular measured construct. For instance, mathematics assessments often produce an overall score as well as scores for geometry, algebra and functions, etc. The production of subscores is justified because, while the sub-components of the assessment are related, items within a sub-content area tend to be related more closely with each other than with items from other sub-content areas. This assumption is very difficult to test for writing because the prompts are only one item—scored multiple times. Subkoviak (1988) described the minimum number of test items on which a score (or subscore) might be based to have even limited reliability to be six items. With only a single item, internal consistency reliability, which is a correlation statistic, cannot be calculated.

Although we cannot calculate a reliability estimate for any of the writing prompt subscores, it would still be expected that the subscores follow a likely pattern. Typically, one would compare the reliability of the subscores with a correlation across subcontent areas to demonstrate that within-component correlations were stronger than between-component correlations. In this case, we can only look at between-component correlations. Those correlations are extremely high. They are typically greater than 0.95 and would be nonsensically greater than 1.00 if we were to correct for unreliability (assuming at least a small amount of measurement error). This phenomenon leads us to conclude that the three component scores may not be discernable measurements from one another. The full test information might be gleaned from a single prompt score and the subscores may not provide useful information for teachers and schools.

One phenomenon that limits how much separation of the writing components we can expect is that scorers tend to score all three prompt components the same. The scores differ by student and by prompt, but if a scorer rates the content component as a 3, that scorer is likely to assign a 3 to structure and conventions as well. As described above, the correlations among the writing prompt components are very high. It is also the case that the scores for the various components are oftentimes exactly the same. It is of course possible that correlations could be high with differing scores (e.g. the correlation between

temperatures measured in Fahrenheit and Celsius degrees would be perfect, but the measurements could be quite different), but that is not the case here. This very high correlation and the tendency for the scores to be the same has led Measured Progress (Kentucky's testing contractor) to estimate parameters for only one component of the writing assessment and then to apply those parameters to the other two components in order to force the parameter estimation software to properly function. Students still receive scores for the other components, but the item parameters are identical for the three components. This process would not have been considered had the scores on the three components differed appreciably.

This phenomenon may be caused by several factors. First, the components may truly not be separable as currently conceptualized. The components may represent separate writing knowledge and skills, but there may not be a sufficient number of students who perform well on one and not the other to differentiate the component skills. Second, this phenomenon may also be an artifact of the scoring process. Currently, scorers receive a response to a prompt and then score that prompt for each of the three components. It is possible that there is some "halo effect" where raters' first impressions of the response influence all three scores. Finally, the scoring process currently uses a 4-point rubric for each component. If rubrics for each component were designed differently, perhaps with additional scoring options, then better differentiation might be achieved.

Because of the concerns outlined above, Kentucky is considering alternate ways of equating and scoring writing. This study examines various options that might be accomplished without substantially altering the test format. The focus of the analyses that follow is to describe how differently students' scores would have been had other equating and scoring procedures been in place. The purpose of these analyses is to illustrate options that might provide some combination of more defensible scores, greater efficiency, and/or stronger year-to-year equating.

Equating Options

Because Kentucky now uses a series of multiple-choice items and writing prompts, and because items of both types are repeated across years, there are several options for year-to-year equating. Best practice would suggest equating using both item types in a common-items equating procedure (such as Stocking-Lord). Using both types of items would provide the strongest link to the past and would account for scoring pattern changes that might occur for one item type, but not the other. This method is used for all other KCCT content area assessments.

The first concern for using both item types for writing comes into play when one considers the manner in which the writing prompts are scored and counted toward a student's final score. There are 24 multiple choice items per test form, but only one writing prompt. The writing prompt is scored on a 0-4 scale by two raters. The two scores are then added, creating a raw score range of 0-8. The process is repeated for each of the three component scores for which the writing prompt is scored. The final result is that the single writing prompt has a potential raw score range from 0-24. So, if the writing prompt

were included in the equating process, we would first need to decide how to treat it. Should a single prompt, for which a single set of item parameters have been created, be treated as one item or three? Should the writing prompt be treated as a 4 point, 8 point, or 24 point item? The weight of the writing prompt for determining a student’s overall score is 50% (24 multiple-choice items each worth 1 point, writing prompt scored on a 4 point scale * 2 raters * 3 components). Should it be given equivalent weight (equal to the multiple-choice items) in the equating solution? These questions, along with dimensionality concerns, have led Kentucky to equate using only multiple-choice items since 2007¹.

In 2009, Measured Progress suggested equating using both item types. Similar to prior years and because of concerns with the very high correlations among the component scores, only a single set of item parameters were estimated for each of the prompts. The plan was to equate treating the prompt as if it measured only one component and had a range of 0-8 raw score points. Because of the questions above and concerns that changing the equating methodology might cause unexpected changes in student scoring patterns, Kentucky decided to continue equating using only multiple-choice items.

The Human Resources Research Organization (HumRRO) was asked to investigate whether adding the writing prompt to the equating set would appreciably impact student scores. Doing so was relatively straightforward. HumRRO simply equated both with and without the writing prompt and calculated the proportions of students scoring in each of the NAPD categories. Results are presented in Table 1.

Table 1. Students Scoring in each NAPD Category Resulting from Equating using Multiple-Choice Items Only Versus Multiple-Choice and Writing Prompt

Performance Category	Grade 5		Grade 8	
	MC Only	MC + WP	MC Only	MC + WP
N	2,908 (5.95%)	3,118 (6.38%)	4,027 (8.28%)	3,471 (7.14%)
A	18,761 (38.40%)	18,145 (37.14%)	24,212 (49.80%)	24,768 (50.95%)
P	22,289 (45.62%)	20,868 (42.17%)	17,998 (37.02%)	18,187 (37.41%)
D	4,899 (10.03%)	6,726 (13.77%)	2,380 (4.9%)	2,191 (4.51%)

Table 1 indicates that the inclusion of the writing prompt, treating it as a single item, results in only small changes in students’ scores. For Grade 8, inclusion never changes the proportion of students in any category by much more than a single percentage point. The change is larger for Grade 5, but still only between 3 and 4

¹ Students choose to respond to 1 of 2 prompts. The impact of choice on equating was not investigated as part of this study.

percentage points where the difference is largest. It is important to note that the number of students impacted by a change in the equating constants is partly dependent on how close a particular raw score is to the scale score cut point. A small change in equating constants might cause a substantial shift in students in a category if a large number of students happen to be on the cusp of scoring in either the next higher or lower performance category. Conversely, if most students' scores are toward the middle of the score distribution for the performance categories, fairly large changes in the equating solution might only cause minor shifts in the proportions of students in any particular category.

Scoring Options

As described above, the correlations among the three components scored from the writing prompt were so high that the parameter estimation software could not estimate parameters for the three components separately. The solution that was agreed upon was to estimate parameters for only the content component. The parameters estimated for content were then used to describe both structure and conventions as well. This allowed for the creation of raw-score-to-scale-score tables and preserved the expected number of raw score points. This solution also allowed Kentucky to provide subscores for all four components of writing. However, there is some concern that the subscores generated do not represent more than a single construct. This might cause schools and teachers to attribute more specificity to students' writing scores than is justified. It might then cause them to adopt inappropriate instructional practices based on spurious indications of strengths and weaknesses of students' written responses.

The assumption was made that it would not matter which component was used to generate the item parameters. This assumption has not been investigated empirically, but seems likely given the high correlations among the components. In fact, because the scores were oftentimes exactly the same across the components, the same result might have been accomplished by simply weighting the single component score by three.

The next set of analyses was designed to investigate a potential improvement in scoring efficiency for the writing prompts. For the reasons described above, Kentucky may not be gaining useful information by scoring each writing prompt in three ways. If the same result could be accomplished from a single score, scorers could focus on either a single component of writing and ignore the other two or they could score the writing prompt more holistically combining aspects of the current scoring methodology to generate a single 0-4 score. To investigate the impact of making such a decision, HumRRO simply eliminated all but one of the prompt subscores, and then generated student scores based on the multiple-choice items and that one component score from the prompt. Those student scores can then be compared directly with operational scores. This process was repeated for each component score to determine if it makes a difference which component is chosen. It should be noted that the prompts were not weighted for this analysis. Each prompt was treated as a 0-8 point item. This reduces the total number of raw score points for the writing assessment from 48 to 32.

The same item response theory software and the same scoring processes that were used for the operational 2009 writing assessment were used for this investigation. The number of and the percentage of students who scored in each performance category (NAPD) based on scoring the prompt for only one writing component was calculated. Table 2 contains the results of the analyses. The “Operational Calibration Sample” column repeats the results from the “MC Only Equating” column in Table 1. These were the operational score percentages for the calibration sample that was used to estimate parameters. They should be directly comparable to the single component scores. The final column indicates the total percentage of students scoring in each category on the operational on-demand writing assessment in 2009 (including any students not in the calibration sample) for additional comparisons.

Table 2. Students Scoring in each NAPD Category Resulting from Estimating Parameters on Different Components

Performance Category	Content Only	Structure Only	Conventions Only	Operational Calibration Sample	Operational State-Level Proportions ²
Grade 5					
N	2,210 (4.52%)	2,116 (4.33%)	1,244 (2.55%)	2,908 (5.95%)	6.16%
A	18,017 (36.88%)	18,562 (37.99%)	14,407 (29.49%)	18,761 (38.40%)	38.47%
P	24,365 (49.87%)	23,655 (48.42%)	27,608 (56.51%)	22,289 (45.62%)	45.38%
D	4,265 (8.73%)	4,524 (9.26%)	5,598 (11.46%)	4,899 (10.03%)	9.99%
Grade 8					
N	3,714 (7.64%)	3,380 (6.95%)	1,960 (4.03%)	4,027 (8.28%)	8.31%
A	23,775 (48.90%)	23,341 (48.01%)	19,457 (40.02%)	24,212 (49.80%)	49.63%
P	18,080 (37.19%)	19,316 (39.73%)	25,613 (52.68%)	17,998 (37.02%)	37.14%
D	3,048 (6.27%)	2,580 (5.31%)	1,587 (3.26%)	2,380 (4.9%)	4.93%

² Downloaded from KDE's web site on 2/10/2010 (<http://www.education.ky.gov/NR/rdonlyres/120033C5-76BD-4D04-AA05-B009AB8ACF37/0/PRS09.xls>)

Table 2 contains a column for “Content Only” as well as an “Operational Calibration Sample” column. These two columns differ, even though content was the only component score for which parameters were estimated during operational calibration and equating. The results differ because the proportions in the “Operational Calibration Sample” column include data from the structure and conventions components, although parameters were not estimated for those components. The “Content Only” column treats the content score as the only score generated from the prompt. The differences in classification were very small between the operational calibration sample and the content-scored assessment. This is not entirely surprising since the item parameters for all component scores for the operational assessment were generated from the content component. Replicating the content parameters to include scores from the structure and conventions components made little difference to the overall test outcomes.

Table 2 also shows that the scores are most different from the operational when the conventions component from the writing prompt is used. This change was substantial despite the high correlations among the components scored using the writing prompt (all > .90). This change indicates that the conventions component, while highly correlated to the other two, may be on a somewhat different scale (e.g. more often higher or lower ratings than the other components, but with essentially the same rank order). This component is treated a bit differently by scorers, which may account for the difference. Student who do not respond to prompts or respond entirely off topic are given 0 points for conventions. If they write toward the topic, they receive at least 2 points. There is no option for receiving 1 point from a scorer. It seems unlikely, given these results, that the content component parameters are interchangeable with the conventions parameters. This issue is most concerning because of the way we treat the components. If we had chosen the conventions component to generate the parameters, students might have scored appreciably differently. However, given the very high correlations, this difference may have been measurement artifact.

Discussion and Conclusions

In 2007, Kentucky shifted the scoring of their on-demand writing assessment from a holistic to an analytic method. Where students previously received only an overall NAPD classification, they now receive a scaled score and analytic/diagnostic information designed to indicate their strengths and weaknesses on four components of writing. Concerns regarding the scaling and equating of the writing assessment, as well as concerns related to the scoring patterns for the writing prompt, led to the investigations presented in this report.

Concerns about the writing prompt have led Kentucky to equate the on-demand writing assessment using only multiple-choice items. An investigation of the impact of that decision was conducted by equating using both multiple-choice items and the writing prompt and making direct comparisons regarding the classification of students. The inclusion of the writing prompt did change the classification outcome for some students. The change was small; typically 1% or fewer students changed classification for Grade 8, and the largest change in classification for Grade 5 was about 3%. These small differences indicate that the decision to omit the prompts from equating does not substantially alter the outcome of the assessment.

However, this study compared multiple-choice equating with equating that included the writing prompt scored for content only. Because the writing prompt is also scored for structure and conventions, it is possible that the equating might have turned out somewhat differently had we chosen one of the other scored components. By mutual decision between Measured Progress, KDE and HumRRO, because the IRT parameter estimation software would not generate plausible parameters for all three highly correlated components, the item parameters for the writing prompt have been generated from the content component only. These were therefore the parameters available for use in equating and for this study. Other evidence from this study indicates that the parameters may have exhibited greater differences had one of the other components been chosen.

The next investigation examined the differences that might have occurred had Kentucky chosen a different component to score on the writing prompt. To examine the potential for differences, each component was used separately to generate parameters. Then, three separate raw-score-to-scale-score tables were produced using the multiple-choice items plus each prompt score component. This had the overall effect of reducing the total raw score from 48 on the operational test to 32 for each prompt component test. The reduction in numbers of score points was expected to cause some differences in the classifications of students. The results of these analyses are in Table 2.

There were some differences in classification depending on the component score that was used. Classifications were very close when results were generated using only the content component or the structure component. However, classification differed the most from operational scores when the conventions component was used. Overall, results suggest that adding two components to students' writing scores without generating parameters to describe those components may create noise in the measurement. Given the high correlations, it is difficult to support the idea that the three components are measuring substantially different content.

Similar writing assessments have been more successful at differentiating among the specified subcontent areas than Kentucky's on-demand assessment. Roid (1994) reported that Oregon was able to generate six separate dimensions of writing, including ideas, organization, voice, word choice, sentence fluency, and conventions. The six dimensions were positively correlated between .49 and .78. Despite these relatively weaker correlations, it was also reported that some of the dimensions could have been combined with little loss of information. Lane's (2006) review of the literature suggests that analytic rubrics typically have the potential to produce distinct information for only a small number (i.e. 2 or 3) of domains.

Recommendations

For equating purposes, Kentucky has chosen to use only multiple-choice items. This report shows that only small differences in categorizations would occur if prompt scores based on content had been included. Should Kentucky desire to include prompt scores for operational equating, it is recommended that revisions to the scoring of the prompts be made first. Using the prompt scores for equating may bolster the accuracy of the equating solution by adding a prompt component to the process component measured by the multiple-choice items. However, the

manner in which the prompt score is included in the equating solution should be deliberate, should not include any replication of item parameters across components, and should be based on the best possible estimate of student achievement on a well-defined identifiable writing construct.

Since Kentucky's on-demand writing assessment produces such highly correlated component scores for the writing prompt, there is not sufficient support to justify continuing to produce student subscores based on these data. The use of a single set of item parameters to represent all three components would not have been justifiable had the content been substantially different. The subscores, at best, provide redundant information and, at worst, add noise to the measurement, potentially promoting poor instructional decisions.

Therefore, we recommended that students be assigned a single score and a single proficiency category for writing barring significant changes to the assessment structure, the scoring methods, and/or the scoring rubrics. If the writing construct is to be divided into multiple components, it is recommended that the decision to do so be informed by indications of convergent/discriminant validity (correlational evidence that different constructs are being measured) and factor analysis (modeling equations that indicate whether multiple factors are being measured by the writing assessment and how those factors are structured). If Kentucky chooses to continue to produce subscores for writing, these analyses will ensure that the subscores represent distinguishable knowledge and skills that can be translated to describe students' strengths and weaknesses among the identified subscores.

Until Kentucky can gather evidence for assessing multiple writing components, it is also recommended that the practice of assigning the same item parameters to all three writing components scored on the prompt be discontinued. Sufficient differences were found when generating distinct parameters for each component and applying them to student scores, calling this practice into question. It is suspected that this difference is because of a scale issue rather than because the students were rank ordered differently by component. Applying the same parameters across components may be adding measurement error to the overall writing scores. Kentucky might consider providing subscores based on the multiple-choice items (a process score) and an amalgam of the currently defined writing prompt components.

References

- Lane, S. Stone, C. A., (2006). Performance Assessment. In R. L. Brennan (Ed.) *Educational Measurement*, pp. 17-64. Westport, CT: Praeger Publishers.
- Roid, G. H. (1994). Patterns of writing skills derived from cluster analysis of direct-writing assessments. *Applied Measurement in Education*, 7(2), 159-170.
- Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement*, 25(1), 47-55.