Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 12, Number 18, December 2007 ISSN 1531-7714

An Investigation of Item Type in a Standards-Based Assessment

Liz Hollingworth, Jonathan J. Beard, and Thomas P. Proctor
University of Iowa

Large-scale state assessment programs use both multiple-choice and open-ended items on tests for accountability purposes. Certainly, there is an intuitive belief among some educators and policy makers that open-ended items measure something different than multiple-choice items. This study examined two item formats in custom-built, standards-based tests of achievement in Reading and Mathematics at grades 3-8. In this paper, we raise questions about the value of including open-ended items, given scoring costs, time constraints, and the higher probability of missing data from test-takers.

The U.S. Department of Education's rules and regulations for the implementation of state assessment systems advocate the use of a variety of item types in state testing programs: "The assessment system must involve multiple approaches with up-to-date measures of student achievement, including measures that assess higher-order thinking skills and understanding of challenging content," (U.S. Department of Education, 2004). In essence, there is an underlying assumption in the federal policy that not only is something substantively different being measured by open-ended items, but the student achievement data yielded are worth the money, time, effort, and introduction of additional scoring error. Given the high-stakes associated with Reading and Mathematics achievement tests used for accountability purposes, we wondered if the item type yielded different information about student achievement in a custom-built test aligned to state standards. States allocate resources for assessment not only to comply with the federal regulations, but also to measure student achievement and school quality. The burden of open-ended item and rubric development would be worthwhile if the actual benefit of additional measurement information were realized: for example if the items were measuring a different and important dimension of the academic construct (e.g., higher order thinking).

This study is an investigation of whether open-ended items provide substantially different information than multiple-choice items on a state-wide, standards-based achievement test that has been written to the specific curriculum standards for the state in which it was used. The data for this study came from a custom-built state assessment, which used the state academic content standards for Reading and Mathematics as test specifications, with both types of items administered simultaneously with an off-the-shelf, multiple-choice, norm-referenced test. Using confirmatory factor analysis to understand the latent traits being tested, we explored student performance in grades 3-8 on both types of items in both subject areas.

BACKGROUND

In the field of educational psychology, much of the literature suggests that item formats should be selected to reflect instructional intent, especially when trying to assess higher level thinking. For instance, Haladyna (1997) writes that open-ended and performance items are more appropriate than selection items "for measuring high-inference mental skills or abilities and some physical skills and abilities where you want the student to construct an answer," (p. 35). Similarly, Marzano and his colleagues at McREL developed a taxonomy they called Dimensions of Learning (Marzano, Pickering, & McTighe, 1993). In order to assess higher order thinking, they argue that performance assessments are a more appropriate item type than selection items because they require students to

construct new knowledge, which is essential to effective learning (p. 26). In addition, Nitko (2004) posits that essay items are valuable because of their unique ability to ask students to explain their choices (p. 181), which in turn gives the evaluator an opportunity to assess higher order learning targets. Multiple-choice items are typically not favored for assessing certain kinds of student learning because of their perceived inability to measure higher order skills. In general, these are the theoretical frameworks that have typically guided the perspectives on item type in the field of educational psychology.

There is a long history of empirical research into the question of item type for achievement tests of Mathematics and Verbal Comprehension in the field of educational measurement. Traub and Fisher (1977) explored the results of different item formats using confirmatory factor analysis and found little evidence of a format effect for Mathematics and weak evidence that the open-ended verbal items measured a different construct. More recent studies from the measurement community have also shown the similarity of assessment data despite changes in item format on the quantitative section of the GRE (Bridgeman, 1992) and on a third grade Reading Comprehension test (van den Bergh, 1990). Empirical evidence of reliability issues with open-ended, constructed-response items comes from research that was conducted using multiple item formats on the Advanced Placement tests (Lukhele, Thissen, & Wainer, 1994; Wainer & Thissen, 1993). When analyzed, the multiple-choice portion of the achievement tests correlated more with the open-ended than the open-ended correlated with itself. One posited explanation for this phenomenon is that it is largely a function of the loss of reliability that comes from the need to score the open-ended items by hand (Dunbar, Koretz, & Hoover, 1991).

Not all measurement research has been conducted in the domains of Mathematics and Reading. For example, Bennett et al. (1991) explored whether two formats assess the same construct in computer science. Like other researchers before them, they found that the open-ended (the authors in this study call it free-response) and multiple-choice items measured analytic thinking in similar ways.

So if tests with different item formats do not measure academic constructs differently, what different kinds of information can be gleaned from various item types? Using an open-ended format for a test of fraction arithmetic with eighth-grade students, Birenbaum and Tatsuoka (1987) suggest that open-ended Mathematics items can give unique diagnostic insight into student misconceptions about the process in the domain. They conceded that the two formats did not measure the construct differently, but that the open-ended items provided the researchers with a unique insight into student thinking. More recently, Briggs, Alonzo, Schwab, and Wilson (2006) have conducted research on the capacity for ordered multiple-choice items to be used for diagnostic purposes when the distracters are built specifically to illuminate common misconceptions students might hold based on pedagogical content knowledge, particularly in Science.

Other scholars have theorized whether some item types bias certain groups of test takers. For example, Webb (1997) argues that multiple-choice tests inherently favor some students over others, so alternative forms of assessment are required to achieve fair measures of student performance (p. 27). In a similar vein, four popular criticisms of objective (i.e. multiple-choice) tests include that they foster a one-right-answer mentality, they narrow the curriculum, they focus on discrete skills, and they under-represent the performance of lower SES examinees (Hambleton & Murphy, 1992). Early research in this area by Rowley (1974) with ninth graders showed that multiple-choice items favored students who were highly test-wise. More recently, research on cognitive demand and item format suggests that different levels of cognition might be tapped depending on question type (Martinez, 1999).

In contrast, other research suggests that performance assessments tap construct-irrelevant factors (Zwick et al., 1993) and open-ended items lend themselves to the introduction of gender bias, since boys and girls respond differently to both visual content and application of knowledge commonly acquired through extracurricular activities (Hamilton, 1998) as well as writing tasks (Beller & Gafni, 2000). What is more, the use of test items that demand verbal abilities for constructs where there is little demand for reading and writing (for instance Mathematics computation or symbolic representation in Physics) can introduce construct-irrelevant variance (Haladyna & Downing, 2004). But when the domain itself is described in terms of writing tasks, as it is for example in essay writing, Ackerman and Smith (1998) argue that asking students to write an essay provides more valid scores than multiple-choice questions.

Rodriguez (2002) summarized the struggle to reconcile the theoretical frameworks of higher order thinking and assessment from the field of educational psychology and the empirical research that has been conducted on item type in the field of educational measurement with the politics of testing and the need for face validity in large-scale assessment programs. He says, "The primary question is: Do multiple-choice (MC) items and constructed-response (CR) items measure the same cognitive behavior? The quick answer is: They do if we write them to do so," (p. 214). In short, he argues that the

item format is not the only characteristic that determines what cognitive constructs are measured in a test.

Practically speaking from a test development perspective, the use of open-ended items increases the chance for additional scoring error. This is because multiple-choice items can be scored electronically, but open-ended items typically require hand scoring by multiple raters to maintain reliable results. This hand scoring is also significantly more expensive than traditional optical scanning methods used with bubble sheets. The estimated cost for using open-ended, performance science items in large-scale testing programs would be "about $34 per class period and $102 per student for a score with reliability of at least 0.80" (Stecher & Klein, 1997). Other concerns include the possibility that language ability might have a confounding effect on the scores for open-ended Social Studies, Science, or Mathematics items and the fact that open-ended items are more likely to be omitted by the examinee than multiple-choice items (Martinez, 1991).

Often, the use of different item types in large-scale state assessment programs for accountability purposes seems to be required mainly for face validity by the U.S. Department of Education. Kane (2006) writes that face validity "refers to the apparent relevance of test tasks to the proposed interpretation or use of scores" (p. 36). It appears that despite the research done in the measurement community about item type since 1977, a face validity stereotype, consistent with the educational psychology literature, persists that says tests with more than one item type yield more valid test scores than tests with only multiple-choice items. In turn, this has affected the way states build their tests for large scale assessment programs.

This study was designed to investigate whether open-ended item types in a standards-based, custom-built state test of Reading and Mathematics are measuring something different from the multiple-choice items.

METHOD

Data sources

In the fall of 2005, 4,111 Ohio students in grades 3-8 answered questions from The Ohio Tests of State Standards (OTSS), a 60-minute augmented, custom-built test in Reading and Mathematics with both open-ended and multiple-choice items that were written to be aligned with the state's academic content standards (see Table 1). The completion criterion (20%) was not met by 198 students, so our analysis was limited to 3,918 students.

The test items were built using the test specifications indicated in the test blueprint from the state of Ohio Department of Education (available online at http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDetail.aspx?Page=3&TopicRelationID=222&Content=31235). Table 2 shows the number of items by type on the OTSS in reading and math. For example, Grade 3 Reading included an open-ended item that required students to complete a graphic organizer table by writing answers to where, when, why, and what questions from a long (351-500 words) sample of informational text. Consistent with the Ohio Department of Education's blueprint, open-ended items on the OTSS were scored using a 0-1-2-3-4 rubric.

Table 1: Ohio State Academic Content Standards in English Language Arts and Mathematics

English Language Arts Standards	Mathematics Standards
Phonemic Awareness, Word Recognition and Fluency	Number, Number Sense and Operations
Acquisition of Vocabulary	Measurement
Reading Process: Concepts of Print, Comprehension Strategies and Self-Monitoring Strategies	Geometry and Spatial Sense
Reading Applications: Informational, Technical and Persuasive Text	Patterns, Functions and Algebra
Reading Applications: Literary Text	Data Analysis and Probability

Table 2: Number of Each Item Type at Each Grade Level on the OTSS

Grade	Reading			Mathematics		
	MC	**OE**	**TOTAL**	**MC**	**OE**	**TOTAL**
3	12	7	19	11	8	19
4	15	7	22	11	8	19
5	17	5	22	12	8	20
6	17	6	23	15	7	22
7	16	7	23	15	7	22
8	15	7	22	15	6	21

In the creation of the open-ended items, we resisted the temptation to write items that could just as easily have appeared as multiple-choice. Items were not given the same stems but different formats, as previous researchers have done, in order to maintain the spirit of building a customized, standards-based test. For instance, in math students were asked to not only compute an answer, but also to show their work because the state standards require that students be able to "model, represent and explain" when computing (Ohio Department of Education, 2005).

The content of the items came directly from the standards for Reading and Mathematics for the state of Ohio. Figure 1 shows a sample item and the scoring rubric for Reading Grade 3. Students were asked to read two passages, one fiction and one non-fiction, about squirrels. Then, in alignment with the Ohio standard, "Create and use graphic organizers, such as Venn diagrams or webs, to demonstrate comprehension," a Venn diagram comparing the two items was presented for the students to write their answers. A second standard, "Compare and contrast information between texts and across subject areas," informed the content of the item itself. As the scoring rubric in Figure 1 indicates, students were scored on their ability to synthesize and compare the information provided in the two reading passages.

Sample

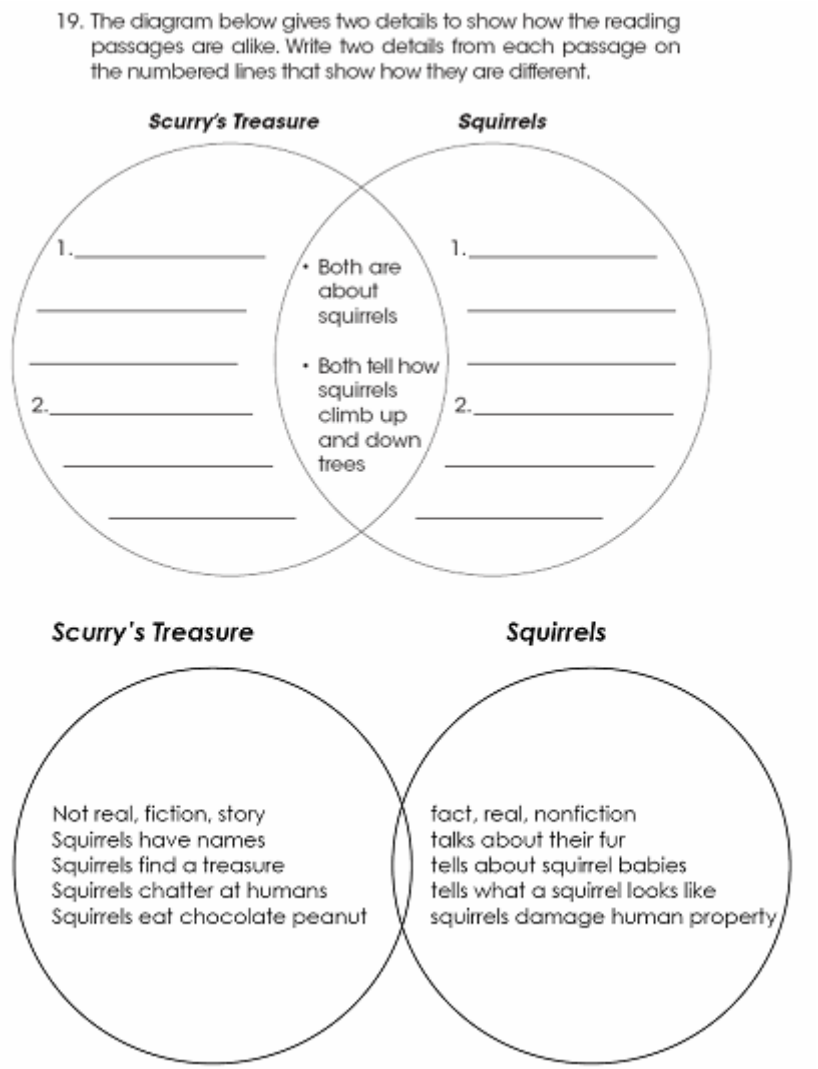
Schools across the state were solicited for participation as part of a field test for augmentation with the Iowa Tests of Basic Skills (Hoover, Dunbar, & Frisbie, 2001). The sample was selected based on school district size, socioeconomic characteristics, and race/ethnicity representation. The number of students at each grade who participated in the study at each grade level can be found in Table 3. The tests were taken at the same time under the same conditions. Because the mixed item format is consistent with the state tests, students would not have been surprised to see both MC and OE items.

After administration, the tests were scored by an independent agency that specializes in hand-scoring of open-ended items. Rubrics were developed by the author team to guide the scorers, who had at least two people score each test to ensure accuracy.

Procedure

In order to combine measurement research about item types with the requirements for large scale assessments, we designed a study to determine whether open-ended (OE) items and multiple-choice (MC) items are related to the same factor. To do this, we conducted a confirmatory factor analysis (CFA) for each test to assess: a) whether a single factor could account for the relationship among OE items, b) whether a single factor could account for the relationship among MC items, c) whether a two-factor model could account for the relationship among OE and MC items, and d) whether the correlation between the two latent factors could plausibly be unity.

The framework provided by classical test theory combined with procedures based upon factor analytic techniques are well-suited to address the research questions of interest. Based upon the tenets of classical test theory (CTT), there are four different sets of assumptions that are used articulate the relationship among true scores: parallel, tau-equivalent, essentially tau-equivalent, and congeneric (Allen & Yen, 2002; Graham, 2006; Gulliksen, 1950; R. Traub, 1994).



Score Point	Definition
4	A **4-point** response fills in four ways the story is different by writing two differences in one circle and two in the other. Examples are listed above.
3	A **3-point** response is one where the student has three of the differences correct, but the fourth one either is not in the story or is left blank.
2	A **2-point** response lists only two differences correctly.
1	A **1- point** response names only one difference.
0	A **0-point** response is blank.

Figure 1: Reading Grade 3 sample item and scoring rubric. Copyright University of Iowa. Used with permission

Table 3: Number of Students at Each Grade Level Who Participated in the Study

Grade	N	Used in Analysis
3	307	209
4	472	467
5	930	920
6	897	858
7	892	854
8	613	610
TOTAL	4111	3918

The analyses in our study were carried out under a congeneric framework because it assumes that each item measures the same attribute, despite the fact that item measurement may be on different scales, with different degrees of precision, and with different amounts of measurement error (Graham, 2006). These assumptions are the most appropriate for two reasons. First, the conception of a true score, as it is defined in classical test theory, could not be considered to be the same for items that are dichotomous when compared to items that are polytomous. We know from the outset that the plausibility of true scores being the same across OE and MC item types is therefore unlikely to hold. Indeed, Qualls (1995) notes that even though different item types may be assessing the same attribute, the differences between items "in a multiformat test can only be modeled through the adoption of a congeneric model for part scores" (p. 113). Second, the generality of the congeneric model assumptions accommodates the primary research question in this study: what is the plausibility of MC and OE items measuring the same construct?

The reliabilities for OE and MC items considered separately were calculated using Raju's general formula for *n* congeneric parts with known lengths (see Table 4). When MC and OE were placed together, reliability was calculated using Raju's formula for two congeneric parts with known test lengths (Feldt & Brennan, 1993).

As a complimentary analysis, Poly-DIMTEST (PD) was used to assess the degree to which these tests exhibited essential unidimensionality (Stout, 1987, 1990). PD is a non-parametric test which investigates the degree to which a single, dominant factor accounts for the responses among dichotomous and polytomous items when other, less dominant factors are present. PD has sufficient power to reject the null hypothesis of essential unidimensionality (Nandakumar, Yu, Li, & Stout, 1998). In our study, PD had unsatisfactory power when the correlation between two abilities was high ($\rho = .7$) and the number of total items on the test was relatively small ($n = 20$). All results for math and for reading from the PD analyses returned non-significant values for Stout's T statistic (see Table 5).

Table 4: Reliabilities of Multiple-Choice, Open-Ended, and Combined Tests for Reading and Math

Grade	Reading			Mathematics		
	MC	OE	Both	MC	OE	Both
3	.53	.71	.73	.55	.36	.66
4	.68	.64	.89	.32	.67	.45
5	.72	.53	1.0	.52	.58	.69
6	.75	.70	1.0	.56	.65	.84
7	.71	.67	.94	.52	.63	.80
8	.72	.69	.88	.64	.58	.94

Table 5: Poly-DIMTEST Results for Reading and Math

	Reading						Mathematics					
Grade	**3**	**4**	**5**	**6**	**7**	**8**	**3**	**4**	**5**	**6**	**7**	**8**
T_L	3.68	5.13	4.67	6.43	6.49	6.33	1.85	2.88	7.31	7.00	8.25	5.12
T_B	2.59	5.97	3.82	6.86	7.84	2.53	2.71	2.11	11.09	8.04	8.95	6.57
T	0.38	-0.32	0.31	-0.16	-0.52	1.24	-.034	0.30	-1.48	-0.40	-0.27	-0.61
p	0.35	0.63	0.38	0.56	0.70	0.11	0.63	0.38	0.93	0.66	0.61	0.73

Model Estimation and Evaluation

We used several CFA models to systematically test our hypotheses about the relationship between the two item types. The various models used here were fit using the means-adjusted weighted least squares algorithm (WLSM) using the Mplus software (Muthén & Muthén, 2004). Parameter estimates were produced using a diagonal weight matrix, and once obtained, a robust asymptotic covariance matrix was used to obtain the standard errors (Flora & Curran, 2004). This method is more appropriate than maximum likelihood estimation (MLE) for analyzing responses that are categorical in nature, and in general, more appropriate for variables that do not meet the multivariate normal assumption (Swygert, 2001).

The various models in this analysis were evaluated based upon goodness of fit criteria. Exact fit and close fit criteria were used to evaluate the plausibility of the proposed models. Exact fit was evaluated using the model χ^2 while close fit was evaluated using the Tucker-Lewis non-normed fit index (TLI) and the root mean squared error of approximation (RMSEA). Although close fit indices are open to interpretation, in this study, values of less than .05 for the RMSEA and values greater than .95 were used for the TLI (Hu & Bentler, 1999). Indices of fit were used conjunctively to assess whether a model was considered well-fitting or not.

Model One: Single Factor CFA for MC and OE Items

Model One states that there is a single factor accounting for the relationship among MC items and that there is a single factor accounting for the relationship among OE items. The model that was fit for each item type is shown in Figure 2. Our null hypothesis was that the construct of interest is unidimensional for each item type and for each subject. So, for each grade, there were four models total: MC for Reading, MC for Math, OE for Reading and OE for Math. Failure to reject the null hypothesis would suggest that each item type within a particular subject

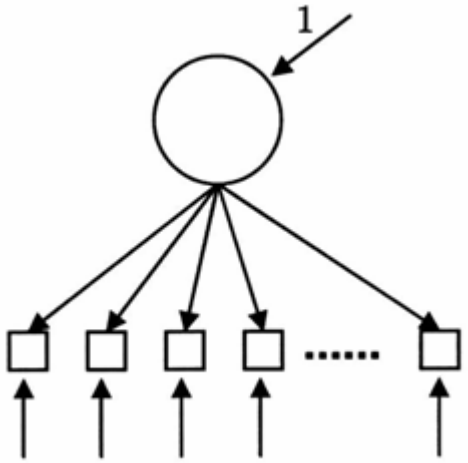


Figure 2: The single factor model for OE and MC items used to assess goodness of fit

is unidimensional. If the null hypothesis is rejected, a single factor model for each item type is not defensible, indicating that a different model might be plausible. If each of the unidimensional models fit, a two factor solution was modeled using MC and OE items within Reading and Math. If either the MC or the OE set of items could not be considered to be unidimensional, the two-factor model would not be presented because defensible evidence exists that a single factor cannot account for the relationships among a set of items. Therefore, a two-factor solution comprised of two unidimensional factors was pursued only when each set of items demonstrated sufficient unidimensionality.

Model Two: Two Factor CFA for MC and OE Items

Model Two states that there is a single factor accounting for the relationship among MC items and that

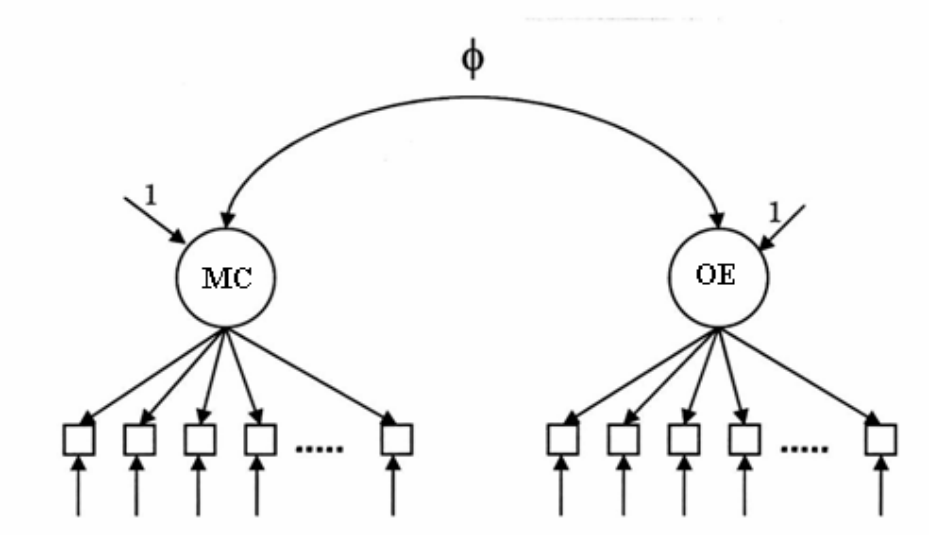


Figure 3: The two factor CFA model used to assess goodness of fit.

there is a single factor accounting for the relationship among OE items, and that these factors correlate. The model that was fit for each item type is shown in Figure 3. Our null hypothesis was that the constructs modeled together would exhibit good fit. So, for each grade, there were two models total: MC and OE for Reading, and MC and OE for Math. Failure to reject the null hypothesis suggests that each item type within a particular subject is correlated. If a two-factor model fits, the correlation between the two factors was tested to determine if the value was significantly different from one.

Model Three: Formal Testing of the Two Factor CFA for MC and OE Items

Model Three is a formal test of the correlation between each of the unidimensional factors. Our null hypothesis was that the latent correlation between OE and MC items is not significantly different from one ($\phi = 1$). We used the MODEL TEST command option in Mplus, which produced a Wald χ^2 statistic. If the phi coefficient was significantly different from one, then a significant chi-square value resulted. If the ϕ coefficient could not be considered significantly different from one, a non-significant χ^2 resulted. Within the assumptions of the congeneric model, retention of the null hypothesis that $\phi = 1$ would suggest that the two unidimensional latent variables measure the same trait. If the null hypothesis is rejected, MC and OE items cannot be described as measuring a common trait.

RESULTS

For Reading, third and fourth grade MC items as well as third, fourth, seventh, and eighth grade OE items exhibited non-significant exact fit χ^2 values (ranging from 109.5 – 15.05). For those grades where the exact fit statistic was significant, close fit statistics indicated acceptable fit (see Tables 6 and 7). Grade 5 OE items exhibited borderline acceptance of fit (χ^2=14.78, df = 9, p = .01, TLI = .96, RMSEA = .05).

For Mathematics, a different pattern emerged. Exact fit indices for fourth grade MC items, as well as fourth and sixth grade OE items exhibited non-significant exact fit χ^2 values. Grade 3 MC items exhibited borderline acceptance of fit (χ^2 = 61.91, df = 44, p = .04, TLI = .90, RMSEA = .04). Grades 5 and 7 for MC items and grade 3 for OE items did not fit.

Based on these results, a single factor solution is not defensible in Mathematics for 3rd grade OE items and 5th and 7th grade for MC items. It is possible that no model at all or a multidimensional model could better account for the relationships among the items, but alternative models were not considered here.

If the two single-factor CFA models were considered plausible, then the two factors were allowed to covary. This produced some estimation problems with some of the tests, and inadmissible solutions resulted. Specifically, the ϕ coefficient was greater than 1.0 in three of the models. However, it was observed that the coefficients were not

Table 6: Analysis for Multiple-Choice Items in Reading and Mathematics

	Reading						Mathematics					
Grade	**3**	**4**	**5**	**6**	**7**	**8**	**3**	**4**	**5***	**6**	**7***	**8**
χ^2	56.55	109.52	164.48	180.71	143.51	113.41	61.91	39.97	178.49	113.54	179.66	106.74
df	54	90	119	119	104	90	44	44	54	90	90	90
p	.38	.08	.00†	.00†	.00†	.05	.04†	.65	.00†	.05	.00†	.11
TLI	.98	.98	.99	.98	.98	.99	.90‡	1.00	.78‡	.97	.85‡	.98
RMSEA	.02	.02	.02	.03	.02	.02	.04	.00	.05	.02	.03	.02
WRMR	.78	.86	.90	.94	.90	.84	.87	.72	1.38	.89	1.13	.84

Note: Grades marked with an asterisk (*) are those grades where a single-factor solution for multiple-choice items is not considered plausible.
† ~ Significant χ^2
‡ ~ TLI does not meet criteria

Table 7: Analysis for Open-Ended Items in Reading and Mathematics

	Reading						Mathematics					
Grade	**3**	**4**	**5**	**6**	**7**	**8**	**3***	**4**	**5**	**6**	**7**	**8**
χ^2	15.24	20.93	14.78	25.69	16.73	15.05	28.47	16.29	35.85	10.84	45.10	18.74
df	14	14	5	9	14	14	9	20	20	14	14	9
p	.36	.10	.01†	.00†	.27	.37	.00†	.70	.02†	.70	.00†	.03†
TLI	1.00	.99	.96	.98	.99	.99	.46‡	1.00	.98	1.00	.96	.97
RMSEA	.02	.03	.05	.05	.02	.01	.10**	.00	.03	.00	.05	.04
WRMR	.50	.60	.67	.73	.52	.50	.86	.49	.74	.43	.89	.67

Note: Grades marked with an asterisk (*) have a single-factor solution for open-ended items and are not considered plausible.
† ~ Significant χ^2
‡ ~ TLI does not meet criteria
** ~ RMSEA exceeds criteria

appreciably larger than one. Since parameter estimation always includes some degree of error, values larger than one were tested along with values that were smaller than one. Thus, grades 3, 5, and 7 Mathematics were excluded from the two-factor solution testing and all other grades were tested for a two factor solution, with the subsequent test of $\phi = 1$.

As shown in Table 8, all of the exact fit statistics for Reading were significant, except for grade 3. However, close fit statistics indicated acceptable fit. The lowest value for the TLI was for grade 8 (.97), and the highest value for the RMSEA was .04, also for grade 8. The ϕ estimates between OE and MC items for reading were also quite high (.96-.99), with correlations greater than one resulting for grades 3 and 7. When the test of unity was conducted on ϕ, the Wald χ^2 was significant for 6th grade ($\chi^2 = 4.27$, df = 1, p = .04) and non-significant for the remaining grades.

Table 8: Analysis of the two factor model combining both item types for Reading

Grade	3*	4	5	6	7*	8
χ^2	178.56	246.59	320.59	395.84	325.09	387.43
df	151	208	208	229	229	208
p	.06	.03†	.00†	.00†	.00†	.00†
TLI	.98	.99	.98	.98	.99	.97
RMSEA	.03	.02	.02	.03	.02	.04
WRMR	.81	.84	.95	.99	.90	1.02
ϕ	1.01	.95	.99	.96	1.00	.96
Wald χ^2	.02	2.95	.20	4.27	.08	1.94
p	.89	.09	.65	.04†	.78	.16

* Inadmissible Solution
† ~ Significant χ^2

Table 9: Analysis of the two factor model combining both item types for Math

Grade	3	4	5	6*	7	8
χ^2	**	200.86	**	317.72	**	207.41
df		151		208		188
p		.00†		.00†		.16
TLI		.96		.97		.99
RMSEA		.03		.03		.01
WRMR		.89		.97		.81
ϕ		.95		1.01		.98
Wald χ^2		.91		.20		.35
p		.34		.65		.55

* Inadmissible Solution
** These models were not estimated because the single-factor solution for either MC or OE items could not be justified.
† ~ Significant χ^2

As shown in Table 9, a two-factor solution for Mathematics was calculated only for grades 4, 6, and 8. The exact fit test was not significant for 8th grade, but it was significant for the remaining grades. However, close fit statistics indicated acceptable fit. The correlations between OE and MC items were also high, and correlations greater than one occurred in grade 6. When the test of unity was conducted on ϕ, the Wald χ^2 was non-significant for all three grades.

DISCUSSION

Our results offer mixed support for multiple-choice and open-ended items measuring the same academic construct. For several grades, an inadmissible solution resulted, indicating that there were some estimation problems that using WLSM could not overcome. When an inadmissible solution did result, the ϕ coefficient was greater than one. None of the error variances in the model was negative. Although the Wald χ^2 statistic indicated that the correlation was not significantly greater than one, substantive interpretations based upon those results were avoided. For the grades that produced an admissible solution, the results were in line with the proposed hypotheses. Even though most of the exact fit statistics indicated that model fit was poor, close fit statistics indicated that the model fit of the two factor solution for MC and OE items was satisfactory.

There was consistency in establishing the unidimensionality of each item type within a subject. Many of the exact and close fit tests demonstrated adequate results for a single factor solution. However, there were exceptions. Grade 3 math OE and grades 5 and 7 math MC tests could not reasonably be considered unidimensional. Estimating ϕ between the latent variables of OE and MC items within a subject area encountered estimation problems for several grades. This left a small number of grades available to investigate whether placing the two unidimensional models in relation to one another would result in a reasonable model. When ϕ was left free to vary, all of the models demonstrated adequate fit. When ϕ was tested to equal unity, the Wald χ^2 test was significant for 6th grade reading, indicating that the correlation between the two latent constructs was significantly different from one, but it was still quite large (.96).

One of the problems with using OE items that often does not appear in the literature about item type is the number of omissions. Some students, when faced with the prospect of writing out an answer, skipped the item altogether. For example, in our Grade 3 sample, omissions on the OTSS for each item ranged from a low of 1.3% in Mathematics to a high of 32% in Reading. That means that as many as one-third of the third grade students provided no data on an OE Reading item. None of the MC items had an omission rate higher than 1%. Other problems less prevalent than omissions that surfaced during scoring were illegible answers and students answering in a language other than English. But without question, the OE item type presented problems of missing data that MC did not. Those who argue on the merits of OE items for their ability to shed light on student thinking should take into consideration that, in fact, some students will be inclined to provide no information about their thinking by simply skipping over the item.

The tacit belief that open-ended items measure higher order thinking skills that multiple-choice items cannot is questionable in policies that guide test development for large-scale programs. Our research suggests that on a test aligned to measure a state's standards, open-ended and multiple-choice items are indeed related to a common factor. Proponents of multiple item types can point to this research and argue that nothing is lost in terms of measurement using open-ended items but that at least using open-ended items removes some of the problems associated with multiple-choice items, for example, guessing. Proponents of multiple choice items could then, in turn, argue that using multiple-choice items gives better control over measurement error, particularly the error that is likely due to raters. We would add that if an argument can be made that information about student learning from both MC items and OE items is of similar quality, then issues of cost, time, efficiency, and reduction in measurement errors should be taken into consideration.

Inclusion of open-ended items has other impacts on assessment beyond issues of measuring the same construct, measuring higher order thinking skills, or measurement error. The U.S. General Accounting Office (2003) compiled a report on the estimated costs of developing, scoring, and reporting assessments required under federal regulations. The GAO estimated that the costs of administration, scoring, and reporting for only machine-scored multiple-choice items would be $1.90 billion, but for states to use machine scored multiple-choice as well as hand-scored open-ended items, the cost would rise to $5.31 billion. The costs for scoring alone were estimated to be $1.23 billion for machine-scorable, multiple-choice items, and for every state to use mixed item formats as $4.59 billion. In conjunction with this research, as well as other research on item format, it appears that the inclusion of open-ended items as they currently are conceived on large scale tests used for NCLB does not yield data which is worth the increased costs, the extended amount of class time needed to test students, nor the time and effort required to score the items accurately.

The policy to include multiple item types in standards-based tests has ramifications of substance for test development with respect to time, money, and reliable scoring. States divert resources for assessment not only to comply with the federal requirements, but also to measure student achievement and school quality. However, because of the additional expenses associated with the development and scoring of open-ended test items, it is critical that the value of the data be well-specified and articulated to policymakers. Certainly, there is an intuitive belief among educators that open-ended items are able to measure something different than multiple-choice items. But further research into the benefits of this kind of data

should be conducted before states are required to spend additional resources on development and scoring of open-ended items in their large-scale state assessment programs.

REFERENCES

Ackerman, T., & Smith, P. (1998). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement*, 12(2), 117-128.

Allen, W. J., & Yen, W. M. (2002). *Introduction to Measurement Theory*. Prospect Heights , IL Waveland Press.

Beller, M., & Gafni, N. (2000). Can item format (multiple choice vs. open-ended) account for gender differences in mathematics achievement? *Sex Roles*, 42(1/2), 1-21.

Birenbaum, M., & Tatsuoka, K. (1987). Open-ended versus multiple-choice response formats—it does make a difference for diagnostic purposes. *Applied Psychological Measurement*, 11(4), 385-395.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29(3)*, 253-271.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*(1), 33-63.

Feldt, L. S., & Brennan, R. L. (1993). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 105-146). New York: American Council on Education/Macmillan.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466-491.

Graham, J. M. (2006). Congeneric and (Essentially) Tau Equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement 66*(6), 930-944.

Gulliksen, H. (1950). *Theory of Mental Tests*. New York: John Wiley & Sons Inc.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Boston, MA: Allyn and Bacon.

Hambleton, R. K., & Murphy, E. (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education, 5*(1), 1-16.

Hamilton, L. S. (1998). Gender Differences on High School Science Achievement Tests: Do Format and Content Matter? *Educational Evaluation and Policy Analysis, 20*(3), 179-195.

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2001). *The Iowa Tests of Basic Skills*. Iowa City, IA: Riverside Publishing.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement, 31*(3), 234-250.

Martinez, M. (1991). A comparison of multiple-choice and constructed figural response items. *Journal of Educational Measurement, 28*(2), 131-145.

Marzano, R. J., Pickering, D., & McTighe, J. (1993). *Assessing student outcomes: Performance assessment using the dimensions of learning model*. Alexandria, VA: Association for supervision and curriculum development.

Muthén, L. K., & Muthén, B. O. (2004). *MPlus* (Version 4.1). Los Angeles: Muthén & Muthén.

Nandakumar, R., Yu, F., Li, H., & Stout, W. (1998). Assessing unidimensionality of polytomous data. *Applied Psychological Measurement, 22*(2), 99-115.

Nitko, A. J. (2004). *Educational assessment of students* (4th ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Ohio Department of Education. (2005). *Academic Content Standards*. Retrieved October 25, 2007, from http://www.ode.state.oh.us/GD/Templates/Pages/ODE/ODEDetail.aspx?page=3&TopicRelationID=333&ContentID=801&Content=32581

Qualls, A. L. (1995). Estimating the Reliability of a Test Containing Multiple Item Formats. *Applied Measurement in Education, 8*(2), 111-120.

Rodriguez, M. C. (2002). Choosing an item format. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 213-231). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Stecher, B. M., & Klein, S. P. (1997). The Cost of Science Performance Assessments in Large-Scale Testing Programs. *Educational Evaluation and Policy Analysis, 19*(1), 1-14.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality

assessment and ability estimation. *Psychometrika, 55*, 293-325.

Swygert, K. A., McLeod, L. D., and Thissen, D. . (2001). Factor Analysis for Items or Testlets Scored in More Than Two Categories. In D. Thissen & H. Wainer (Eds.), *Test Scoring*. Mahwah, New Jersey: Lawrence Earlbaum Associates.

Traub, R. (1994). *MMSS Reliability for the Social Sciences: Theory and Applications*. Newbury Park, CT: Sage Publications.

Traub, R., & Fisher, C. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement, 1*(3), 355-369.

U.S. Department of Education. (April 28, 2004). *Standards and Assessments Peer Review Guidance.* Retrieved October 17, 2006, from http://www.ed.gov/policy/elsec/guid/saaprguidance.pdf

United States General Accounting Office. (2003). *Title I: Characteristics of tests will influence expenses; Information sharing may help States realize efficiencies.* Retrieved 12/2/06, from http://www.gao.gov/new.items/d03389.pdf

van den Bergh, H. (1990). On the construct validity of multiple-choice items for Reading comprehension. *Applied Psychological Measurement, 14*(1), 1-12.

Wainer, H., & Thissen, D. (1993). Combining Multiple-Choice and Constructed-Response Test Scores: Toward a Marxist Theory of Test Construction. *Applied Measurement in Education, 6*(2), 103-119.

Webb, N. (1997). Criteria for alignment of expectations and assessments in mathematics and science education. Research Monograph No 6. Madison: University of Wisconsin-Madison, National Institute for Science Education.

Note

The views, opinions, and positions expressed in this paper are those of the authors and do not necessarily reflect the views, opinions, or positions of the Ohio State Department of Education. The Ohio Tests of State Standards are not in any way associated to or endorsed by the Ohio State Department of Education.

Citation

Hollingworth, Liz, Beard, Jonathan J., & Proctor, Thomas P. (2007). An Investigation of Item Type in a Standards-Based Assessment. *Practical Assessment Research & Evaluation*, 12(18). Available online: http://pareonline.net/getvn.asp?v=12&n=18

Correspondence

Please address correspondence

Liz Hollingworth
Assistant Professor
University of Iowa
College of Education
340C Lindquist Center
Iowa City, IA 52242
319-335-5409 (voice)
319-335-6038 (fax)

liz-hollingworth [at] uiowa.edu