A Multitrait-Multimethod Validity Investigation of the 2002 Massachusetts Comprehensive Assessment System Tests\textsuperscript{1,2}

Dean Goodman

University of Massachusetts Amherst

April 30, 2004

\textsuperscript{1} Center for Educational Assessment MCAS Validity Report Number 5. (CEA-531). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

\textsuperscript{2} The research was completed as part of a contract between the Massachusetts Department of Education and the University of Massachusetts Amherst. However, the Department of Education is not responsible for any errors that remain in the paper nor should it be assumed that they agree with the findings and conclusions.
Each year, students from across Massachusetts are required to take part in Massachusetts Comprehensive Assessment System (MCAS) tests. These tests assess academic achievement in multiple subject areas (e.g., mathematics, English language arts [ELA], and history and social science) across multiple grades (e.g., grades 3, 4, 6, 7, 8, and 10). MCAS tests provide key measures of student performance in relation to state academic standards, and are used at the tenth grade as a statewide graduation requirement for all students in Massachusetts public schools. While a body of evidence has been compiled that illustrates the meaningfulness of MCAS test scores (see Massachusetts Department of Education, 1999), continued efforts to demonstrate the validity of interpretations based on MCAS test scores are required. This paper outlines a validation effort to evaluate the constructs measured by MCAS Mathematics and ELA tests administered in the early and late stages of schooling (grades 4 and 10).

A Framework for the Study: Construct Validation Through the Multitrait-Multimethod Matrix

In the current Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999), all test scores are viewed as measures of some construct (p. 174). A construct is a theoretical variable that must be inferred from multiple types of evidence. Two general types of evidence that can be used to assess the degree to which test scores measure a particular construct are convergent and discriminant validity evidence. Convergent validity evidence can be obtained by examining “the relationship between test scores and other measures of the same construct” (p. 174). Discriminant validity evidence can be gathered by examining “the relationship between test scores and measures of different constructs” (p. 175).
Convergent and discriminant validity evidence can be examined systematically using the multitrait-multimethod (MTMM) matrix approach proposed by Campbell and Fiske (1959). This approach involves correlating test scores with other measures of the trait (i.e., construct) purportedly measured by the test (i.e., monotrait correlations), as well as with measures of different traits (i.e., heterotrait correlations). Correlations involving similar measurement methods (i.e., monomethod correlations) and different measurement methods (heteromethod correlations) are also considered in the MTMM matrix approach, as are the reliabilities of each measure. The joint consideration of all these correlations allows construct-relevant trait variance to be distinguished from construct-irrelevant method variance (p. 102).

Campbell and Fiske (1959) outlined four criteria for evaluating a MTMM correlation matrix. As noted by Campbell and Fiske (1959) and Pitoniak, Sireci, & Luecht (2002), the first criterion offers evidence of convergent validity and the second, third, and fourth criteria offer evidence of discriminant validity. The four criteria for evaluating a MTMM correlation matrix are as follows:

1. Correlations for the same trait measured with different methods (referred to as “convergent validities”) should be significantly different from zero and “sufficiently large to encourage further examination of validity” (Campbell & Fiske, 1959, p. 83);

2. Each convergent validity (same trait measured with different methods) should be higher than the other correlations found in its row and column that measure different traits by different methods;
3. Convergent validities (same trait measured with different methods) should be higher than correlations among different traits measured with the same method; and

4. The pattern of correlations among traits should be the same both within a method and across methods. (Pitoniak et al., 2002, p. 499)

Purpose of the Study

The purpose of this study is to investigate the construct validity of scores from 2002 MCAS tests using the MTMM approach. Campbell and Fiske’s (1959) criteria are used to evaluate convergent and discriminant validity evidence obtained across two tests designed to measure distinct constructs (mathematics and ELA) using two different measurement methods (multiple-choice and open-response questions).

Method

The MCAS tests and data used in this study are described in this section, followed by the procedure used to construct the relevant MTMM matrices.

Selection and Descriptions of the MCAS Tests

In 2002, MCAS tests were administered across six grades (grades 3, 4, 6, 7, 8, and 10) and three subject areas (mathematics, ELA, and history and social science). Students in grades 4, 8, and 10 were administered two different tests (mathematics and ELA in grades 4 and 10, and mathematics and history and social science in grade 8). Students in grade 3 took the ELA Reading test. All MCAS tests contained at least two types of items: multiple-choice items (scored 0, 1) and open-response items (scored 0, 1, 2, 3, 4). Mathematics tests also contained short-answer items (scored 0, 1) and ELA tests also
The MTMM approach requires test scores representing at least two traits, each of which were measured by at least two different methods, across a common group of people. In this study, fourth- and tenth-grade MCAS scores were selected for validation because they satisfied the data requirements of the MTMM approach, and measured the same pair of traits (mathematics and ELA). Because of concerns about the amount of written text associated with items on MCAS mathematics tests and the possibility that reading ability impacts mathematics test results, a MTMM analysis of mathematics and ELA scores was also considered to be of more practical and theoretical interest than a comparable analysis of mathematics and History and Social Sciences test scores (the only other set of MCAS scores that could be subjected to a MTMM analysis).

The grade 4 Mathematics test consisted of 39 scoring items: 29 multiple-choice (MC) items, 5 short-answer (SA) items, and 5 open-response (OR) items. The grade 10 Mathematics tests consisted of 42 scoring items: 32 MC items, 4 SA items, and 6 OR items. The grades 4 and 10 ELA tests consisted of 42 scoring items: 36 MC items, 4 OR items, and one extended writing prompt (WP). Consistent with the requirements of the MTMM approach, only test scores based on measurement methods common to all tests (i.e., MC and OR items) were considered in this study.

Data

Scored item response data were obtained for the 2002 MCAS grades 4 and 10 Mathematics and ELA tests. The grade 4 datasets consisted of approximately 21,800 examinees; the grade 10 datasets consisted of approximately 63,700 examinees. The
mathematics and ELA datasets for each grade were merged using SPSS so that scored responses for both tests were matched to common student identification numbers. Students who did not have responses to either of the MC or OR components of the tests were deleted from the datasets. These procedures resulted in merged datasets of N=21,557 for grade 4 and N=62,570 for grade 10. To facilitate cross-grade comparisons, equivalently-sized datasets were created by randomly selecting 20,000 examinees from each dataset. All of the MTMM analyses carried out in this study were conducted on these random samples of 20,000 examinees.

Construction of the MTMM Matrices

For each test, a total raw score across MC items and a total raw score across OR items were calculated for each examinee. While these scores are not reported on MCAS, they provide a convenient way to evaluate test score variance across different measurement methods and are well-suited for MTMM analyses.

Reliability estimates were calculated for the MC and OR components of each test using coefficient alpha. Within each grade, Pearson product-moment correlations between each test score were also calculated. To address possible concerns about the unreliability of the measures and to better evaluate the true correlations among the constructs, a second set of correlations corrected for attenuation was calculated (i.e., disattenuated \( r_{xy} = r_{xy} / \sqrt{r_{xx} r_{yy}} \)). Consistent with the recommendations of Messick (1989, p. 47), these disattenuated correlations would act as the primary focus in the construct validation process.
Results

The MTMM matrices obtained through the correlation are presented in Figures 1 to 4. The uncorrected correlations for the grades 4 and 10 MCAS tests are presented in Figures 1 and 2, respectively; comparable correlations corrected for attenuation are presented in Figures 3 and 4. In all cases, the uncorrected correlations are much lower than the relevant correlations corrected for attenuation. Nonetheless, within each year, the pattern of results obtained is consistent across the uncorrected and disattenuated correlation matrices.

Within each MTMM matrix, reliabilities for each test component appear on the main diagonal in italicized font within parentheses. Convergent validities (monotrait-heteromethod correlations) are presented as bolded text within brackets. Correlations for different traits using the same measurement method (i.e., heterotrait-monomethod correlations) are surrounded by solid lines. Correlations for different traits using different measurement methods (i.e., heterotrait-heteromethod correlations) are surrounded by dashed lines. Results for the disattenuated grades 4 and 8 matrices (Figures 3 and 4) are summarized below in relation to Campbell and Fiske’s criteria for evaluating MTMM matrices.

1. Monotrait-heteromethod correlations

In both grades, the convergent validities (i.e., correlations for the same trait measured by different methods) are large and significantly different from zero ($p<.01$). Similar convergent validities are obtained in grades 4 and 10. In both grades 4 and 10, the highest convergent validities are obtained for the mathematics trait ($r = .92$). While still considered large, smaller convergent validities are obtained for the ELA trait.
Figure 1. Uncorrected multitrait-multimethod correlation matrix for 2002 grade 4 MCAS Mathematics and ELA tests.

<table>
<thead>
<tr>
<th>Method</th>
<th>Multiple-Choice Questions</th>
<th>Open-Response Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Trait</td>
<td>Math</td>
</tr>
<tr>
<td>Multiple-Choice Questions</td>
<td>Math</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ELA</td>
<td>0.71</td>
</tr>
<tr>
<td>Open-Response Questions</td>
<td>Math</td>
<td>[0.74]</td>
</tr>
<tr>
<td></td>
<td>ELA</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Note. $N = 20,000$. All correlations are significant at the .01 level. Italicized values in parentheses are reliabilities; bolded values in brackets are convergent validities. Heterotrait-monomethod groupings are enclosed by solid lines; heterotrait-heteromethod groupings are enclosed by dashed lines.

Figure 2. Uncorrected multitrait-multimethod correlation matrix for 2002 grade 10 MCAS Mathematics and ELA tests.

<table>
<thead>
<tr>
<th>Method</th>
<th>Multiple-Choice Questions</th>
<th>Open-Response Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Trait</td>
<td>Math</td>
</tr>
<tr>
<td>Multiple-Choice Questions</td>
<td>Math</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ELA</td>
<td>0.64</td>
</tr>
<tr>
<td>Open-Response Questions</td>
<td>Math</td>
<td>[0.78]</td>
</tr>
<tr>
<td></td>
<td>ELA</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Note. $N = 20,000$. All correlations are significant at the .01 level. Italicized values in parentheses are reliabilities; bolded values in brackets are convergent validities. Heterotrait-monomethod groupings are enclosed by solid lines; heterotrait-heteromethod groupings are enclosed by dashed lines.
Figure 3. Disattenuated multitrait-multimethod correlation matrix for 2002 grade 4 MCAS Mathematics and ELA tests.

<table>
<thead>
<tr>
<th>Method</th>
<th>Trait</th>
<th>Multiple-Choice Questions</th>
<th>Open-Response Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Math</td>
<td>ELA</td>
</tr>
<tr>
<td>Multiple-Choice Questions</td>
<td>Math</td>
<td>(0.84)</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>ELA</td>
<td></td>
<td>(0.84)</td>
</tr>
<tr>
<td>Open-Response Questions</td>
<td>Math</td>
<td>[0.92]</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>ELA</td>
<td>0.69</td>
<td>[0.79]</td>
</tr>
</tbody>
</table>

Note. $N = 20,000$. Italicized values in parentheses are reliabilities; bolded values in brackets are convergent validities. Heterotrait-monomethod groupings are enclosed by solid lines; heterotrait-heteromethod groupings are enclosed by dashed lines.

Figure 4. Disattenuated multitrait-multimethod correlation matrix for 2002 grade 10 MCAS Mathematics and ELA tests.

<table>
<thead>
<tr>
<th>Method</th>
<th>Trait</th>
<th>Multiple-Choice Questions</th>
<th>Open-Response Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Math</td>
<td>ELA</td>
</tr>
<tr>
<td>Multiple-Choice Questions</td>
<td>Math</td>
<td>(0.85)</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>ELA</td>
<td></td>
<td>(0.84)</td>
</tr>
<tr>
<td>Open-Response Questions</td>
<td>Math</td>
<td>[0.92]</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>ELA</td>
<td>0.62</td>
<td>[0.80]</td>
</tr>
</tbody>
</table>

Note. $N = 20,000$. Italicized values in parentheses are reliabilities; bolded values in brackets are convergent validities. Heterotrait-monomethod groupings are enclosed by solid lines; heterotrait-heteromethod groupings are enclosed by dashed lines.
Given these results, the first criterion that the convergent validities are significantly different from zero and sufficiently large to encourage further examination of validity is satisfied for the grades 4 and 10 MCAS tests.

2. Heterotrait-heteromethod correlations

In both grades, the convergent validities for mathematics are larger than the other correlations found in the row and column that measure different traits by different methods (i.e., heterotrait-heteromethod correlations). Thus, these convergent validities satisfy this second criterion proposed by Campbell and Fiske. In both grades, the convergent validities for ELA (.79 in grade 4 and .80 in grade 10) are much larger than the heterotrait-heteromethod correlations obtained for mathematics MC items and ELA OR items (.69 in grade 4 and .62 in grade 10); however, these convergent validities are slightly smaller than the heterotrait-heteromethod correlations obtained for mathematics OR items and ELA MC items (.81 in both grades). Based on these results, the grade 4 and grade 10 MCAS tests did not fully meet the second criterion proposed by Campbell and Fiske.

3. Heterotrait-monomethod correlations

The third criterion outlined by Campbell and Fiske is that convergent validities should be higher than correlations among different traits measured with the same method (i.e., heterotrait-monomethod correlations). In grade 4, the convergent validities (.92 for mathematics and .79 for ELA) are higher than the correlation between MC items measuring mathematics and ELA traits (.84) and the correlation between OR items that measure the two traits (.77). Slightly larger differences exist in grade 10, with the convergent validities of .92 for mathematics and .80 for ELA being higher than the
correlation between MC items measuring mathematics and ELA traits (.76) and the
correlation between OR items that measure the two traits (.74). Thus, the grade 4 and 10
MCAS tests fully meet Campbell and Fiske’s third criterion.

4. Patterns of heterotrait correlations

Campbell and Fiske’s final criterion for evaluating MTMM matrices is that the
pattern of correlations among traits should be equivalent both within a method and across
methods (i.e., within both the heterotrait-monomethod correlations and the heterotrait-
heteromethod correlations). Unfortunately, the application of this criterion requires test
scores representing more than two traits or methods of measurement. Since only two
general traits (mathematics and ELA) and two measurement methods (MC and OR
questions) are reviewed in this study, the grades 4 and 10 MCAS tests could not be
evaluated in terms of this fourth criterion.

Discussion

This study represents an initial application of the MTMM construct validation
approach on MCAS test data. The purpose of this study was to use this approach to
gather convergent and discriminant validity evidence for investigating the construct
validity of scores from 2002 MCAS tests. In general, the findings of this study help
confirm the construct validity of scores from the fourth- and tenth-grade Mathematics and
ELA tests. The findings of this study also raise several issues that MCAS test developers
might consider when developing and validating future MCAS tests.

In this study, significant convergent validity evidence is obtained for the fourth-
and tenth-grade Mathematics and ELA tests. In both grades, MC and OR questions
designed to measure the same construct are highly correlated (ranging from .79 to .84). It
is interesting to note that after disattenuation these convergent validities are almost identical across both grades (.84 for mathematics in both grades, .79 for ELA in grade 4 and .80 for ELA in grade 10). This provides some evidence about the consistency with which each construct is measured using MC and OR items across early and late stages of public schooling.

The MTMM analyses provide some useful discriminant validity evidence related to the fourth- and tenth-grade MCAS tests. While generally favorable, some of this evidence suggests additional work is required to fully demonstrate the construct validity of scores of MCAS tests.

As is desired, the convergent validities obtained in both grades for mathematics are larger than the correlations of different traits across different methods (i.e., heterotrait-heteromethod correlations). Unfortunately, in both grades the convergent validities for ELA are larger than only one of the two heterotrait-heteromethod correlations. The high heterotrait-heteromethod correlation of .81 obtained in both grades for mathematics-OR/ELA-MC is a source of concern, and warrants further attention by the MCAS test developers. The findings of this study might, for example, indicate the influence of a reading-related construct on mathematics OR items (OR items on the mathematics test required students to answer a series of questions and, as a result, typically have more text associated with them than MC items [see Massachusetts Department of Education, 2002]). Efforts by the developers of MCAS to improve the correlation between MC and OR items on the ELA tests, or to reduce the correlation between mathematics OR items and ELA MC items would help improve the construct validity of the MCAS scores. For example, given the very high convergent validities
obtained for mathematics (.92, when corrected for attenuation), it might be useful for test
developers and those responsible for overseeing the scoring of ELA tests to talk to their
counterparts in mathematics for strategies for improving the correlation between MC and
OR test items. To minimize the correlation between mathematics OR items and ELA
MC items, the mathematics test developers might, for example, try to reduce the amount
of text associated with OR mathematics items.

Additional discriminant validity evidence was provided by comparing the
convergent validities obtained at each grade with correlations among different traits
measured with the same method (i.e., heterotrait-monomethod correlations). Consistent
with the third criterion proposed by Campbell and Fiske, all convergent validities were
higher than the relevant heterotrait-monomethod correlations. Still, the relatively high
heterotrait-monomethod correlations obtained across grades (particularly the .84
correlation obtained for the MC items in grade 4) may indicate that method variance may
be a factor in MCAS tests, and should be an issue of consideration for the developers of
future MCAS tests (e.g., test developers may wish to explore ways to minimize the
impact that item format has on the results of different tests).

Unfortunately, a final source of discriminant evidence, evaluating the pattern of
correlations among traits within both the heterotrait-monomethod groupings and the
heterotrait-heteromethod groupings, could not be obtained from the current study. This
represents a key limitation of using the MTMM approach for construct validation of
existing MCAS scores, and possible remedies should be considered.

To evaluate patterns of correlations among traits within and across methods,
additional methods and traits need to be included in the MTMM matrices. The inclusion
of additional methods, such as short answer items to ELA tests, would impact other features of MCAS tests (e.g., the ability to accurately compare the results of current and past tests), and so is not advised. The need for additional traits in the MTMM analyses, however, will likely be addressed in the next several years as a result of the No Child Left Behind Act (NCLB). To comply with NCLB, Massachusetts will need to assess student progress in reading and mathematics in each of grades 3 through 8 and at least once during grades 10 through 12, and in science at least once during grades 3 through 5, grades 6 through 9, and grades 10 through 12. Once the same group of students are assessed in three different subject areas, a complete set of discriminant validity evidence can be compiled. This will facilitate a more comprehensive application of the MTMM approach than was possible in the current study.

The NCLB requirement that all states report diagnostic information on NCLB-related assessments emphasizes a further need for additional, more comprehensive validation efforts concerning MCAS test scores. If subdomain scores are to be reported on MCAS tests, it is critical that construct validation studies be carried out on these scores. Doing so would help address important requirements outlined in the Standards for Educational and Psychological Testing, specifically the need to demonstrate the distinctiveness of the separate scores and to show that the interrelationships of those scores are consistent with the construct(s) being assessed (AERA et al., 1999, p. 20). Consequently, additional and regular applications of the MTMM procedures to MCAS test scores appear necessary.
Conclusion

The current study illustrates the application of a useful procedure for evaluating the construct validity of scores from MCAS tests. In general, the findings of this study help confirm the construct validity of the MCAS scores from fourth- and tenth-grade Mathematics and ELA tests. However, a number of findings, such as the high correlation between OR mathematics items and some ELA items, are of some concern and warrant further attention. Further research is recommended to better demonstrate the construct validity of MCAS test scores, especially as new types of MCAS tests and test scores are developed and reported.
References


