Vertical Scaling for the MCAS\textsuperscript{1,2}

Craig Wells and Lisa Keller

University of Massachusetts Amherst

August 30, 2005

\textsuperscript{1} Center for Educational Assessment Validity Report No. 13. (CEA-584). Amherst, MA: University of Massachusetts, Center for Educational Assessment. The authors are pleased to acknowledge the editorial contributions of Ronald Hambleton.

\textsuperscript{2} This work was carried out under a contract between the University of Massachusetts Center for Educational Assessment and Measured Progress, Dover, New Hampshire, and the Massachusetts Department of Education.
Vertical Scaling for the MCAS

Craig Wells and Lisa Keller
University of Massachusetts Amherst

Introduction

Thanks in part to the No Child Left Behind Act of 2001, measuring academic progress over time (i.e., grade levels) has become an important element of state-assessment programs. The development of a common scale across grade levels of a given content area, known as vertical scaling, provides states the capability of tracking achievement growth over grade levels. However, the implementation of vertical scaling is controversial due to the many technical and conceptual problems encountered (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999). The current paper examines several issues related to developing a vertical scale for the MCAS math and English language arts (ELA) tests. First, a general vertical scaling design will be proposed. Second, the feasibility of establishing a vertical scale for the MCAS mathematics and ELA tests will be addressed in light of the proposed design.

Vertical Scaling Design

A common-item design used in conjunction with item response theory (IRT) may be used to provide a vertical scale for grades 3, 4, 5, 6, 7, 8, and 10 of the MCAS math or ELA. (Please note that in this paper, the word “common” refers to the items used for linking the grade score scales together. We understand that with the MCAS program, “common” refer to the operational portion of each assessment, and therefore our use of the word “common” may be initially confusing.) In the common-item design, adjacent grade levels share identical (i.e., “linking” or “common”) items, in addition to unique items administered only to those examinees at a grade level, and are used to link the scales. The common items allow the scales to be linked so that growth over grade levels may be assessed. An example of a common-item design for the MCAS math or ELA tests is shown in Figure 1.
Figure 1. Common-item design for MCAS math or ELA tests

<table>
<thead>
<tr>
<th>Grade</th>
<th>Item Block</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
</tr>
<tr>
<td>5</td>
<td>C</td>
</tr>
<tr>
<td>6</td>
<td>D</td>
</tr>
<tr>
<td>7</td>
<td>E</td>
</tr>
<tr>
<td>8</td>
<td>F</td>
</tr>
<tr>
<td>10</td>
<td>G</td>
</tr>
</tbody>
</table>

Note: Not shown are the item blocks defining the unique items administered within each grade level.

The common-item design shown in Figure 1 indicates that each test shares identical items with adjacent grade levels. For example, grades 3 and 4 share item block B while item block C is shared between grades 4 and 5. A grade-by-grade chained linking procedure is used to construct a vertical scale from the lowest to highest grade by first selecting a base grade (e.g., grade 3). Then, the common item blocks between adjacent grade levels are used to link all of the scales onto the scale of the base level. The details of this procedure are provided below. Criteria for choosing common items are discussed in a later section and in considerable detail by Kolen and Brennan (2004).

**Implementing the Common-item Design**

The goal of the following procedure is to place the item parameter estimates and ability distribution of each grade onto the scale of the base grade level, which in this case was chosen to be grade 3. A good topic for discussion is the choice of grade to establish the scale (i.e., the starting point), but such a discussion will not be addressed in this paper.

1. At each of the seven grades, estimate the item parameters for the relevant Item Response Theory (IRT) model(s) via the marginal maximum likelihood estimation (MMLE) algorithm and ability distributions via expected-a-posteriori (EAP) estimates, separately.
2. Linking grades 3 and 4:
a. Select the common items between grades 3 and 4 that do not exhibit differential item functioning (DIF); these items will be used as the anchor in linking grades 3 and 4.

b. Place the item parameter estimates and ability distribution for grade 4 onto the scale of grade 3 using the non-DIF common items identified in step (2a) via the test characteristic curve (TCC) method (see Stocking & Lord, 1983).

3. Linking grades 4 and 5:
   a. Select the common items between grades 3 and 4 that do not exhibit DIF; these items will be used as the anchor in linking grades 4 and 5.
   b. Place the item parameter estimates and ability distribution for grade 5 onto the scale of grade 4 using the non-DIF common items identified in step (3a) via the TCC method.

4. Continue the process to link the remaining grade levels so that grade 6 is linked to grade 5, grade 7 to grade 6, grade 8 to grade 7, and finally, grade 10 is linked to grade 8.

A few additional comments are informative to supplement the above description. First, because the inclusion of DIF items in the anchor set almost certainly will have a detrimental impact on the linking of adjacent grade levels (Shepard, Camilli, & Williams, 1984), it is crucial to remove any DIF items from the anchor set before estimating the linking coefficients. In this case, an iterative DIF detection technique (Kim & Cohen, 1992) may be applied using the DIF statistic of choice to purify the anchor set.

Second, the IRT model(s) used to define the latent ability metric must be appropriate in that their assumptions should be met. One such assumption is that the model accurately reflects the underlying item response function (IRF); i.e., the model fits the data (Hambleton, Swaminathan & Rogers, 1991). Therefore, the vertical scaling procedure, when implemented using IRT, requires that the IRT model fit the data in order for the scale to be valid across grades. One particular consequence of using a poorly fitting model is that the item may exhibit DIF when the two groups have disparate ability distributions, as observed between grade levels (Bolt, 2002). As a result, it is important to exclude items from the linking procedure for which the model does not fit.
Selection and Placement of the Common Items

The common items between adjacent grade levels ultimately define the scale for which achievement growth will be judged. Therefore, the selection of common items for a particular item block requires careful consideration so as to increase the chances of having high quality linking items (e.g., avoiding items that exhibit unintended dimensionality). First, the selection of the particular common items between adjacent grade levels depends primarily on the appropriateness of the item’s content with respect to the curriculum taught and/or skill set required at the particular grade levels. For instance, it is crucial that the linking items are appropriate for the grade level for which they are administered in that the examinees have been taught the material recently or that the skill required to answer the item is routinely used in the curriculum. When inappropriate items are administered to examinees, several consequences are possible, not the least of which are the negative consequences on test validity. One such unfortunate event occurs when examinees from a lower grade actually outperform examinees from an upper grade. This may happen when, for example, mathematical material in 8th grade is not taught in 10th grade or the skill is not used daily.

Second, it is optimal to balance the number of items chosen from the lower and upper grade level so as to minimize any deleterious consequences associated with an unbalanced anchor item set in the common item linking design shown in Figure 1. For instance, selecting all the items from a lower grade will likely result in an unbalanced item block with respect to test specification and item difficulty. Moreover, having a well-balanced test with respect to item difficulty (i.e., the difficulty parameters are evenly distributed) increases the chances that the equity property in equating will be met (Baker, 1984).

Once the items have been selected, it is important to consider each linking item’s location on the test so as to minimize any context effects (i.e., context effects occur when an item parameter’s value depends on its test location). Therefore, it is best to place each item at approximately the same position in the respective tests. One possible solution is to place all of the linking items at the end of the test. However, if the common items are placed at the end of the test, it is prudent that the test not be speeded (a speeded test has time pressure, so some examinees won’t get all the way through the test). The examinees
for whom the test is speeded (who omit many of the questions near the end of the test) must be removed from the linking process using a mixture modeling approach (Bolt, Cohen, & Wollack, 2002; Yamamoto & Everson, 1997).

Assessing Grade-to-grade Achievement Growth

Once a vertical scale has been established, there are three score scale properties that are useful in describing grade level development. The first property is grade-to-grade achievement growth and is typically measured by the difference between the mean of the ability distribution for the adjacent grades. The second property is grade-to-grade variability, which may be displayed using the within grade standard deviation of the ability for adjacent grades. The last property, separation of grade distributions, may be illustrated by graphing the cumulative distribution function for adjacent grade levels. The following effect size, proposed by Yen (1986), may be used to summarize the separation of grade distributions

\[
\tilde{\eta} = \frac{\hat{\mu}(\theta)_{\text{upper}} - \hat{\mu}(\theta)_{\text{lower}}}{\sqrt{\hat{\sigma}^2(\theta)_{\text{upper}} + \hat{\sigma}^2(\theta)_{\text{lower}}}/2},
\]

where \(\hat{\mu}(\theta)_{\text{upper}}\) and \(\hat{\sigma}^2(\theta)_{\text{upper}}\) is the mean and variance of the upper adjacent grade level; and \(\hat{\mu}(\theta)_{\text{lower}}\) and \(\hat{\sigma}^2(\theta)_{\text{lower}}\) is the mean and variance of the lower adjacent grade level. Essentially, the effect size standardizes the grade-to-grade mean differences in achievement.

The Feasibility of Vertical Scaling for the MCAS

The feasibility of vertical scaling for the MCAS depends primarily on whether the intended unidimensional scale may be established across grade levels for math and ELA. Two of the major assumptions of vertical scaling are that the tests are assessing comparable content across the grade levels and that similar dimension(s) are being assessed in each grade. Meeting these two assumptions is critical for the success of vertical scaling. Severely violating these assumptions will likely result in misinterpreting achievement growth over grade levels. In fact, it is possible to show no growth over several grades due to a poorly constructed vertical scale even when growth is present.
Of all of the subject areas in education, vertical scaling appears most appropriate in mathematics and reading/ELA since these two subjects are taught continuously throughout a student’s schooling (this is in contrast to areas such as science where the content changes from grade-to-grade). Unfortunately, math and reading/ELA may not be summarized by one or even a few shared dimensions across all of the desired grade levels, especially if the curriculum is distinct in each grade (a vertical scale cannot capture important grade-specific dimensions). Therefore, the more unique a curriculum is from grade to grade, the less likely a vertical scale will be appropriate. As a result, a feasible vertical scale for math or ELA requires consistency in instructional content or curriculum across grades and/or uniformity of rehearsal of skill and subskills being measured. Determining the consistency of the dimensionality of tests in a subject across grade levels is a very good research topic and involves a combination of judgmental and statistical evidence.

Math vs. ELA

As described above, vertical scaling is more problematic in domains that are highly curricular dependent from grade to grade. Therefore, domains that measure constructs that rely more heavily on curriculum are less appropriate for vertical scaling (Lloyd & Plake, 1987). As a result, we will argue that a vertical scale would be more problematic to establish for math compared to ELA, though we recognize the importance of supporting this statement with empirical evidence.

Theory in the field of education has been better able to grasp the complicated nature of learning to read, and subsequently, reading to learn, particularly compared to the scarcity of consensus regarding instructing mathematical skills. Theories for reading instruction have been better validated and have a more specific and well-defined skill hierarchy, which is reflected in the consistency of the curriculum standards over early grade levels (National Research Panel, 2000). The essential enabling skills for proficient readers - phonemic awareness, alphabetic awareness, phonics and fluency - have been specifically delineated in the Massachusetts frameworks in K to 3 (Massachusetts Department of Education, 1997). These skills are innately rehearsed while children are learning new, more advanced skills along the reading hierarchy after grade 3. Once these
skills are mastered, subsequent instruction focuses on complex vocabulary and comprehension strategies. As a result, the instructional focus shifts from teaching students to read to teaching students to learn through reading. Establishing a vertical scale in math is less feasible due to the lack of consensus on a specific skill hierarchy beyond math fact fluency (Massachusetts Department of Education, 1996). The breadth of mathematics curriculum, and goals of math instruction, are vast and diffuse. In addition, practice of the most basic skills, such as addition and subtraction, are not consistently rehearsed compared to the enabling skills for reading.

However, even though establishing a vertical scale in ELA may be more feasible compared to math, it still may not be reasonable considering that there will exist some unique dimensionality within grade levels that will not be captured by the vertical scale, resulting in a misleading assessment of achievement growth.

Evidence from Other States

Several major test publishers have successfully been able to report achievement test scores on a vertical scale (e.g., TerraNova-CTB/McGraw-Hill, California Achievement Tests-CTB/McGraw-Hill, Comprehensive Test of Basic Skills-CTB/McGraw-Hill, Iowa Test of Basic Skills, Stanford Achievement Test-Harcourt Educational Measurement, and Metropolitan Achievement Test-Harcourt Educational Measurement). The technical aspects of each test (e.g., equating design) are diverse indicating that there is more than one reasonable method that may be used to establish a vertical scale. However, the success of each of these tests is partly due to the fact that they are not highly curricular dependent, unlike state assessments.

A few state assessment programs have attempted to establish vertical scales (e.g., Florida and Mississippi). However, the technical and conceptual difficulties of creating a continuous scale over grade levels has left other states searching for alternative methods of assessing achievement growth mandated by No Child Left Behind. For example, South Carolina has implemented a form of vertically moderated standards that attempts to measure growth but avoid the pitfalls of vertical scales thus far.
References


