Validity Issues Involved in Cross-Grade Statements About NAEP Results

David Thissen
University of North Carolina, Chapel Hill

January 2012
Commissioned by the NAEP Validity Studies (NVS) Panel

George W. Bohrnstedt, Panel Chair
Frances B. Stancavage, Project Director

The NAEP Validity Studies Panel was formed by the American Institutes for Research under contract with the National Center for Education Statistics. Points of view or opinions expressed in this paper do not necessarily represent the official positions of the U.S. Department of Education or the American Institutes for Research.

The NAEP Validity Studies (NVS) Panel was formed in 1995 to provide a technical review of NAEP plans and products and to identify technical concerns and promising techniques worthy of further study and research. The members of the panel have been charged with writing focused studies and issue papers on the most salient of the identified issues.

Panel Members:

Albert E. Beaton
Boston College

Peter Behuniak
University of Connecticut

George W. Bohrnstedt
American Institutes for Research

James R. Chromy
Research Triangle Institute

Phil Daro
University of California, Berkeley

Lizanne DeStefano
University of Illinois

Richard P. Durán
University of California, Santa Barbara

David Grissmer
University of Virginia

Larry Hedges
Northwestern University

Gerunda Hughes
Howard University

Robert Linn
University of Colorado at Boulder

Ina V.S. Mullis
Boston College

Scott Norton
Louisiana Department of Education

Gary Phillips
American Institutes for Research

Lorrie Shepard
University of Colorado at Boulder

David Thissen
University of North Carolina, Chapel Hill

Karen Wixson
University of North Carolina, Greensboro

Project Director:

Frances B. Stancavage
American Institutes for Research

Project Officer:

Janis Brown
National Center for Education Statistics

For Information:

NAEP Validity Studies (NVS)
American Institutes for Research
2800 Campus Drive, Suite 200
San Mateo, CA 94403
Phone: 650/ 843-8100
Fax: 650/ 843-8200

Acknowledgments

Thanks for helpful comments and suggestions from Al Beaton, Gary Phillips, Peggy Carr, Andy Kolstad, Fran Stancavage, George Bohrnstedt, Deborah Holtzman, Larry Hedges, Lorrie Shepard, Ina Mullis, Peter Behuniak, Gerunda Hughes, Jeff Nellhaus, Jim Chromy, Andreas Oranje, and Gloria Dion. Any misstatements are, of course, my own.

CONTENTS

The reading and mathematics measures of the National Assessment of Educational Progress (NAEP) have been, and continue to be, reported on scales that appear to have the properties of "cross-grade" scales: Reported scores are higher for 8th-grade students than for 4th-grade students, and higher for 12th-grade students than for 8th-grade students.[1,2] Historically, these scales were developed in ways that were intended to support cross-grade interpretation; however, the degree of support for and endorsement of such interpretations has varied over the past two decades. Nevertheless, these score scales invite interpretive statements about the results that can be divided into two categories, each requiring support from different kinds of validity evidence:

Statement (1) "One year's growth is (approximately) x NAEP scale points"

An example of the use of this kind of score interpretation appeared in a blog called "The Daily Howler" by Bob Somerby on April 7, 2010 (at http://www.dailyhowler.com/dh040710.shtml), in the context of a commentary about a *Washington Post* editorial on the 2009 NAEP reading results. The blog included the following:

> In 4th-grade reading, American kids seem to have shown good progress since 1998.… Since 1998, white kids have gained 5 points on the NAEP scale; by the rough rule of thumb which is often used, this would be equivalent to roughly one-half year of growth.… Black 4th-graders have gained 12 points in reading during that period, roughly 1.2 years. Hispanic kids have made the same gain—12 points, 1.2 years. Warning! This "rough rule of thumb" is very rough; we long for the day when some major newspaper asks NAEP officials to discuss the meaning of these score gains in some serious detail.… But this rough rule of thumb has been widely used; its surface logic is apparent. (Don't ask.) If we do apply that rough rule of thumb, those score gains seem quite consequential.
>
> By the way: Children scoring at the 10th percentile have also gained 12 points in reading during that period… This suggests that our current lowest achieving 4th-graders are more than a year ahead of their counterparts from 1998. If that's true, it's remarkable progress.
>
> The picture in 8th-grade reading is worse.… Since 1998, white 8th-graders have only advanced 3 points on the NAEP scale in reading—perhaps three-tenths of a year. Black 8th-graders have advanced only 2 points. That said, Hispanic kids have advanced 6 points—theoretically, more than half a year. Kids at the 10th percentile have also advanced by only 3 points…

[1] Due to changes in the 2005 mathematics framework that differentially affected grade 12, the main NAEP 12th-grade mathematics assessment is not currently reported on the same "cross-grade" scale as grades 4 and 8.
[2] The geography and U.S. history measures have also been reported on cross-grade scales, as has science (historically, from 1986 until the scale was replaced in 1996). However, due to the higher salience and frequency of administration of the reading and mathematics assessments, this essay will concentrate on the latter.

Somerby notes in his "warning" that the "rule of thumb" (10 points per year, for the NAEP reading scale) used throughout his interpretation of score change is not officially sanctioned. Nevertheless, this description of the results clearly places them in a context that could make the scores more comprehensible for many who want to interpret NAEP results.

In addition, in the current climate in which policymakers may seek to set goals for educational improvement in the metric of assessment results, it would be informative for such policymakers if they knew whether a 5-point gain on the NAEP reading scale, for example, represented something like academic progress for one month, or for one year. A goal that suggests students should make an extra month or two of progress in an academic year might be considered much more reasonable than a goal suggesting two years' progress in a single year.

Statement (2) "Subgroup A in grade X performs approximately the same as subgroup B in grade Y"

In a presentation to the Graduate Record Examinations (GRE) Board on October 15, 2010, entitled "Addressing Achievement Gaps: A Leading Role for the GRE Board," Michael T. Nettles showed graphics using NAEP's long-term trend data to point out that (on the NAEP score scale for the 2004 long-term trend mathematics assessment) "Black and Hispanic 17-year-old students achieve at the level of White 13-year-olds."[3] This conclusion is straightforward to draw from the standard NAEP graphics that place the long-term trend results for 9-, 13-, and 17-year-olds, or the main NAEP results for 4th-, 8th-, and 12th-grade students, on the same graph with a common vertical axis. Kolstad (2004) points out that NAEP graphics invite this interpretation, but that official NAEP reports no longer make such interpretations in the text associated with the graphics.

Statement (2a) "A score of NNN in grade X has the same meaning as a score of NNN in grade Y"

Interpretive statements that fit the model shown in (2a) are less-embellished versions of interpretations that fit the model shown in (2): Instead of reference to two groups that have the same average score, such interpretations refer directly to the score. It will become clear that the counterclaim—that score *NNN* represents different performance in grade *X* than it does in grade *Y*—is both an argument that interpretations of form (2a) are false, and the primary attack on the validity of claims of form (2).

There is, of course, only a blurry distinction between interpretive statements of form (1) and those of form (2/2a). If one knows the value for "one year's growth" (say, 7 points) for a particular grade transition (say, from grade 7 to grade 8), and one observes a subgroup that has an average score 21 points below the overall average, then it is a short, if not perfectly accurate, leap to say that group is "three years

[3] While it happens that this example uses results from the NAEP long-term trend, similar descriptions are also given of main NAEP results.

behind." The difference in accuracy or validity between interpretations of the first and second kinds involves the span of years: Interpretations across a four-year span are subject to more threats to validity than interpretations about a single year.

The Goals of This Paper

This paper ultimately seeks to make two points: (1) Different evidence is needed to support the two categories of interpretive statements described above, and (2) Insufficient evidence is currently available to support either category of interpretations for NAEP. Either further research is required to support either or both of these classes of interpretations, or greater clarity is needed in the presentation of NAEP results to discourage such interpretations. The conclusion of this essay will be that evidence can and should be assembled to support, and make more precise, interpretations of the first kind ("one year's growth"), while interpretations of the second kind (cross-group comparisons across four-year spans) should be discouraged.

Background

NAEP's cross-grade scales have a checkered history. Cross-grade scales were a feature of the "new design" that brought item response theory (IRT) scaling to NAEP. The reading cross-grade scale was first established in 1984 (Beaton, 1987). Moving forward, the entire cross-grade 1986 NAEP reading scale was linked back to the 1984 scale. At the same time, in 1986, new cross-grade scales were constructed for mathematics and science. 1988 NAEP reading was linked to the 1984–1986 reading scale (Beaton, 1988). 1990 NAEP appears to have been complicated: 1990 reading was linked to both 1984 and 1988, and new multidimensional, cross-sectional, cross-grade scales were created for mathematics and science. (Unidimensional, composite mathematics and science scales also were constructed and linked back to 1986.) (Johnson & Allen, 1992).[4] The reading scale in current use dates back to 1992 (Johnson & Carlson, 1994).

Subsequently, a decision was made by the National Assessment Governing Board (NAGB) that "within-age scales should be used whenever feasible" (Haertel, 1991, p. 15). Haertel's (1991) *Report on TRP [Technical Review Panel] Analyses of Issues Concerning Within-Age versus Cross-Age Scales for the National Assessment of Educational Progress* provided a set of arguments for that decision. Excerpts from that report will be used liberally in subsequent parts of this document because the arguments have not changed over the intervening years.

After 1992, NAEP scale maintenance remained within grade, although continued use of the originally cross-grade scales within grade produced the appearance of cross-grade scales without any real checks on their validity. That is, the current main NAEP mathematics scale was developed in 1990, and the current main NAEP reading scale was developed in 1992. Each was analyzed across grades only in the

[4] The science assessment was subsequently placed on a new within-grade scale in 1996 (Kolstad, 2004).

base year. Starting in 2005, grade 12 mathematics was uncoupled from the cross-grade scale, and it has been reported on a separate scale with a different metric since then.

Until recently, cross-grade blocks of items appeared on both the reading and mathematics assessments. In the period from 2003 to 2009, cross-grade blocks on the mathematics assessment were gradually released without replacement. The last cross-grade blocks for mathematics were administered in 2009; the 2011 mathematics assessment uses completely nonoverlapping sets of items across grades. Meanwhile, while the number of cross-grade blocks decreased slowly on the reading assessment until 2007, that number was subsequently increased in anticipation of the development of a new cross-grade scale based on the new 2009 reading framework and data from the 2009 administration.

In 2004, NAGB adopted a *Resolution on the NAEP 2009 Reading Framework* (http://www.nagb.org/what-we-do/resolution-09.htm) that stated, in part, "The 2009 NAEP reading assessment will establish a new trend line… Achievement will be reported on an overall cross-grade scale, allowing NAEP to show the development of reading skills through the years of schooling as well as reporting trends over time." In October 2004, a "Technical Panel Meeting to Discuss the Implementation of Within- and Cross-Grade Scaling for the NAEP 2009 Reading Assessment" was held (Wise & Hoffman, 2004). That meeting included presentations on, and discussions of, a number of issues involved with reinstitution of a cross-grade scale for the reading assessment.

Between 2004 and 2010, Educational Testing Service (ETS) performed a number of retrospective studies, followed by analyses of the 2009 operational data, and concluded that the 2009 NAEP reading assessment could be linked to the original 1992 scale. Relatively little information is publicly available about those studies, with the exception being a set of PowerPoint slides (McClellan, Donoghue, Gladkova, & Xu, 2005) from a presentation at the conference on "Longitudinal Modeling of Student Achievement" that was held at the Maryland Assessment Research Center for Education Success in 2005, and the following statement on the NAEP website (http://nces.ed.gov/nationsreportcard/reading/interpret-results.asp):

> Average reading scale score results are based on the NAEP reading scale… The composite reading scale is defined differently in the 2009 framework than in the previous reading framework, but special analyses determined that the 2009 results could be compared to those from previous years.… The results for all three grades are placed together on one reporting scale. In the base year of the trend line, the three grades are analyzed together to create a cross-grade scale. In subsequent years, the data from each grade level are analyzed separately and then linked to the original cross-grade scale established in the base year. Comparisons of overall student performance across grade levels on a cross-grade scale are acceptable; however, other types of comparisons or inferences may not be supported by the available information. Note that while the scale is cross-grade, the skills tested and the material on the test increase in complexity and difficulty at each higher grade

level, so different things are measured at the different grades even though a progression is implied."

Validity Issues

Fundamentally, cross-grade scales represent linked tests. As such, the interpretation that is most demanding of validity evidence is of the form of statements (2) or (2a) above. Examples might be "Black and Hispanic 12th-grade students achieve at the level of White 8th-grade students" or "A NAEP reading score of 267 has the same meaning in grades 8 and 12." If either of these statements were taken completely literally, it would require evidence that, if the grade 8 assessment had been administered to Black and Hispanic 12th-grade students, the students would have obtained the same average score as they did on the 12th-grade assessment; or that 8th-grade students who obtained scores around 267 on the grade 8 assessment would also receive scores around 267 on the grade 12 assessment. However, the statement about the reading scale that appears on the NAEP website (quoted immediately above) shies away from this very strong interpretation, concluding with the disclaimer that: "… while the scale is cross-grade, the skills tested and the material on the test increase in complexity and difficulty at each higher grade level, so different things are measured at the different grades even though a progression is implied."

On the other hand, statements of the form of (1) above, "One year's growth is (approximately) *x* NAEP scale points," are potentially much less demanding of validity evidence. To obtain appropriate evidence, "all" that is required is the administration of the same test at adjacent grades, and computation of the empirical value of "one year's growth" as the difference between the average scale scores. The "all" is in quotes because data collection to assemble this evidence would be very expensive in a complex national survey such as NAEP, and it would be made even more expensive if questions were raised about conditioning "one year's growth" on demographic characteristics or score levels.

Historically, much of the discussion about the validity of cross-grade scales for NAEP has been from an "all or nothing" perspective: If one truly believes that one has a well-constructed cross-grade scale that measures a single construct (that is, a scale that is unidimensional)—or, as is the case of NAEP, a scale that is a fixed composite of several unidimensional scales—then interpretations of both the first and second kinds are automatically valid. In much of the historical discussion of cross-grade scales in NAEP, the opposite has been taken to be true as well: If one disbelieves in any feature of the cross-grade scale, then neither kind of interpretation is valid.

A point of this paper is to make the case that different validity evidence is needed for statements of forms (1) and (2) above, and further that it would probably be useful to obtain evidence to support interpretations of form (1) (one year's growth), while it may be infeasible to collect evidence to support interpretations of form (2) (subgroups across grades).

Commentary on Objections to Interpretations Based on Cross-Grade Scales

Haertel's (1991) Report on TRP Analyses of Issues Concerning Within-Age Versus Cross-Age Scales for the National Assessment of Educational Progress ably summarizes most objections that have been raised to interpretations of test scores based on cross-grade scales. These objections have not really changed over the years, so Haertel's report is used to organize this section. It is worth noting at this point, however, that Haertel bases his critique on an assumption that cross-grade scales must have much stronger psychometric properties than those attributed to NAEP cross-grade scales by the statement currently on the NAEP website. Haertel (1991, p. 2) says the following:

> In particular, a score on a cross-age scale, say 300, should represent the same overall level of proficiency—and the same mix of skills—for a 9-year-old or a 13-year-old or a 17-year-old. That level of attainment by a younger versus an older child would probably be interpreted differently, of course. A proficiency level considered excellent for a 4th-grader might be barely adequate for a 13-year-old. Nonetheless, if there is a common scale, then a given score on that scale should carry some definite implication as to what the child earning it knows or can do.

By contrast, the NAEP website says (about the reading scale) that "the skills tested and the material on the test increase in complexity and difficulty at each higher grade level," which is very different from Haertel's "a score … should represent the same overall level of proficiency—and the same mix of skills—for a 9-year-old or a 13-year-old or a 17-year-old." It is also worth noting that, wherever possible, Haertel's (1991) document drew illustrations from the mathematics assessment, for which there are more clear differences between the content of the items across ages/grades than for the reading assessment.

More than half of Haertel's (1991) report is devoted to summarizing a large number of analyses checking the internal consistency of the original NAEP cross-grade scales. The report concludes, "*If* one accepts the reasonableness of cross-age scales, then the ETS *implementation* of cross-age scaling procedures for the 1990 mathematics assessment appears satisfactory, as do the results of selected examinations of the 1986 and 1988 NAEP mathematics and reading data" (emphasis in the original).

Since 1990 there have been advances in statistical methodology that can be used in cross-grade scaling. For example, at the 2004 technical panel meeting to discuss the implementation of within- and cross-grade scaling for the NAEP 2009 reading assessment, Patz (2004) described ways that the now-standard NAEP scaling methodology could be augmented with conditioning variables for grade and for interactions between grade and demographic groups, to provide direct unbiased estimates of the kinds of cross-grade group differences that appear in interpretations of kind (2), above. Reckase (2004) presented approaches for the use of multidimensional models that may be more realistic for NAEP data. Yen (2004) described various data-collection designs that could be used to develop cross-grade

scales and cautioned that issues of multidimensionality of the assessment should be examined carefully. Nevertheless, an overview is that the arguments are not really about the statistical technologies that are used to create these scales; the arguments are about how much they cost, what they mean, and whether NAEP would be better off with or without them.

Disbelief in Cross-Grade Scales

Haertel (1991) describes several challenges to the validity of "quantitative" interpretations based on the cross-grade scale. The first example Haertel offers is of form (1) (one year's growth), which he calls "Interpretations in terms of 'grade equivalents.'" Haertel (1991, p. 12) writes,

> This interpretation, along with several others, depends critically on the linearity of the cross-age scale. Perhaps the best illustration was given by the 1986 reading anomaly, wherein a change of about 3 percent in the probability of a 17-year-old's answering a reading exercise correctly was translated to a drop of "a full grade level" in 17-year-old reading proficiency between 1984 and 1986. This figure was reached by taking the difference in overall mean scale scores for 13-year-olds and 17-year-olds, treating this as the gain to be expected over four years, and dividing by four to define expected annual growth. The "grade level" metric made a very small absolute change in performance appear much more substantial. Because 13-year-olds and 17-year-olds are typically tested on different content, very strong assumptions are entailed in expressing the difference between 1984 17-year-olds' performance (on grade 12 reading) and 1986 17-year-olds' performance (on grade 12 reading) in terms of the difference in scale scores corresponding to 13-year-old performance (on grade 8 reading) and 17-year-olds' performance (on grade 12 reading).

There are (at least) two levels of interpretation of Haertel's concern about "very strong assumptions." A minimalist interpretation is that there is no reason to assume that growth in reading proficiency is linear between grades 8 and 12; indeed, vertical scales constructed on a year-to-year basis involve decelerating growth curves. In a presentation at a recent NAEP Design Summit, Tirre and Oranje (2010) tabulated results from eight nationally normed reading tests with cross-grade scales. All showed decelerating growth, with one year's growth equal to a (decreasing) fraction of the within-grade standard deviation for grades 2–12. Similar results are routinely observed for assessments of mathematics.

In this interpretation, the only thing wrong with the interpretation expressed in the reporting of the reading anomaly is that one-fourth the difference between the grade 8 mean and the grade 12 mean is the wrong value for "expected annual growth." Indeed, in this interpretation, the value is too small; if the growth curve is decelerating, one fourth of that value is *more* than "a grade level" at the top (grade 12). It is this interpretation that could be answered with data collected at shorter intervals, like one or two years apart instead of four.

Somerby's blog commentary used this same kind of "rule of thumb" (a rounded, divided-by-four value of the four-year change in average reading scores on the NAEP scale) as "one year's growth." Given the expected decelerating nature of growth on assessment scales commonly used to measure academic achievement, that is clearly wrong. However, one can turn the question around and ask "how wrong can it be?" The answer to that would be "not very wrong." Assuming only that NAEP scores would grow more or less like those on any other academic achievement test, we know that "one year's growth" would be a (decreasing) fraction of the within-grade standard deviation, which for NAEP scores is in the low to mid 30s. Ten points would be one-third of that, which is very near the average for "one year's growth" for reading in grades 3–8 tabulated by Bloom, Hill, Black, and Lipsey (2008) and reported by Tirre and Oranje (2010). Bloom et al. (2008) also report values for mathematics tests that are about one -third larger than for reading. So if NAEP is basically like all other achievement tests, we know that "one year's growth" is between about 8 points (growth of 0.25 within-group standard deviations), which is about right for reading around grade 8, and 17 points, which is about right for mathematics around grade 4, based on the data presented by Bloom et al. (2008). Ten points for reading may not be "right," but it cannot be far wrong.

However, this minimalist interpretation was probably not what Haertel was referring to with "very strong assumptions." This point is illustrated with the next example, "Comparisons of growth rates" (Haertel, 1991, pp. 12–13):

> On page 55 of *The State of Mathematics Achievement* the statement appears,
>
> > *As would be expected, 12th-graders had higher average proficiency than did 8th-graders, who in turn performed better than 4th-graders. Eighth-graders performed, on average, 50 points higher on the scale than did 4th-graders. The 12th-graders, however, on average, performed only 30 points higher on the scale than did the 8th-graders.*
>
> This statement at least implicitly suggests that growth in mathematics proficiency is more rapid between grade 4 and grade 8 than between grade 8 and grade 12. No further interpretation is offered, but the reader's attention is directed to the scale point descriptions, which characterize performance at levels 200, 250, 300, and 350. Inspection of the scale point descriptions highlights the fragility of any "equal interval" interpretations for the NAEP proficiency scale. In what sense is the distance from the 200 description to the 250 description the same as the distance from the 250 description to the 300 description, for example? In fact, it is very difficult to say *anything* useful about the fact that 8th-graders outperform 4th-graders by more points than 12th-graders outperform 8th-graders.

In this excerpt, Haertel expresses disbelief in *any* interval-scale interpretation of the cross-grade scale, suggesting that a finer-grained analysis, that divides the 50-point gain from grade 4 to grade 8 into gains of 16, 13, 11, and 10 points per year for those four years, and the 30-point gain from grades 8 to 12 into 9, 8, 7, and 6 points per year, would not be satisfying.

It is difficult to see an interval scale in verbal performance descriptions; for the scale, we must turn to our statistical models. The IRT models we use assume that achievement is a latent, unobservable construct. As a result, we have no way of knowing whether our observable measures are isomorphic with the latent variable or not. There are theoretical (that is, nonempirical) reasons to use shapes like logistics for the trace lines. The consequence of such a choice, however, is that we get a shape for the growth curve that is, in the case of achievement test data, (empirically) invariably decelerating.

Haertel (1991) also expresses concern about interpretations that compare "gaps" across grades using the vertical scale, or that compare high performers at one grade with low performers at another, but these are just more reflections of the disbelief expressed in the *Report* regarding any interval interpretation of the IRT scale.

Disbelief in anything like an interval scale interpretation of cross-grade scales is not unique to either Haertel's *Report* or to commentary written twenty years ago. Derek Briggs' 2010 presentation at the International Meeting of the Psychometric Society was based on the strong position that such scales are no more than ordinal.

This leads to a question: Setting aside Platonic argument about the interval nature of IRT score scales, is it desirable, or would it be useful, to behave in a theory-agnostic way and use special studies to determine the size of "one year's growth" on the NAEP scale by testing students one year apart, while eschewing any use of the scale to make comparisons across grades 4–8–12? This could be done by administering the same assessment in grades 3, 4, and 5 and in grades 7, 8, and 9, for example. The sole purpose of this exercise would be to make more precise the value—in points on the NAEP scale—of "one year's growth," as an aid in interpreting changes in scores over time or score differences between groups within one grade. It is not even necessary to have a cross-grade scale to do this. Nor is it necessary to believe the score scale has interval properties; one only needs to take at face value the conventional measurement of achievement with test scores.

Questions About the Relation of Cross-Grade Scales and the Curriculum

Haertel (1991) also describes a second class of "curricular" interpretations of cross-grade scales, as follows:

> *Linear conception of the curriculum*. Cross-age scaling may encourage a view of the curriculum and learning in terms of progress along simple, unidimensional continua spanning (at least) grades 4 through 12. Such a view tends to support the idea that advanced, higher-order skills must be reserved for the later years of schooling, and children during their earlier years need to concentrate on largely meaningless, decontextualized "tool" skills in preparation for that later application. An alternative conception (and scaling) of curriculum *within* grade levels can direct attention to higher-level application and problem solving for younger children as well as older, and

> can provide assessment information more in keeping with current reform initiatives in various curriculum areas.

Without careful analysis of the degree to which the frameworks represent "learning progressions" across grades, it is very difficult to separate fact from rhetoric in that argument. There may be some of both.

A second argument Haertel makes against the use of cross-grade scales is specific to the context of modern (i.e., post-1990) NAEP:

> *Anchor point descriptions*. The use of anchor point descriptions intended to apply equally to 4th, 8th, and 12th grade students at a given proficiency level is problematical for the same reason as the use of cross-age proficiency scales. The location of different skills and abilities is largely determined by the conventions of curriculum organization, so that children at a given grade level are *necessarily* confined to a relatively narrow scale score region, and consequently, a very limited number of anchor point descriptions. Within-grade scales with separate anchor point descriptions for 4th, 8th, and 12th grade students would depict more clearly the range of achievement levels and variety of attainment patterns characterizing different subgroups within each grade level, and would not divert attention to largely meaningless comparisons between the knowledge and skills of children four or eight years apart in age.

The question is, to what extent are comparisons of children four or eight years apart in age "largely meaningless"?

In her presentation at the conference *Linking and Aligning Scores and Scales* at Princeton in June of 2005, Wendy Yen (2007) used the following parable to illustrate the fragility of interpretation of cross-grade scales:

> I have been interested in vertical scales for a bit more than 25 years. When I was about five years old, I used to follow my father around as he did home improvements. He had a folding ruler with which I would play. It was yellow, with hinged 1-foot lengths that would unfold (making a nice thwacking sound) to 6 feet. If I held the extended ruler at one end, it would curve gracefully through space. To my disappointment, if I leaned it too much to the side, one of the lower hinges would suddenly bend sharply.
>
> A vertical scale is akin to a folding ruler. Although educational achievement tests tend to have very strong first factors, they are multidimensional, paralleling changes in the curriculum. This dimensionality changes both within and across test levels. The direction of the scale (i.e., the relative importance of the different dimensions) changes as the test levels become more difficult. Thus, the scale bends or curves through space. Connections between some levels are stronger (i.e., have tighter hinges) than others, and sometimes the links between levels are too loose to maintain a sturdy connection between test levels.

A folding ruler measures very well over spans over which it is held straight. But how wide are those spans for NAEP's cross-grade scales?

Yen's metaphor makes it easy to explain why "one year's growth" may be an interesting concept, while comparisons across four years may be of less interest. One year's growth is like one segment of the folding ruler: straight and linear. Across several segments (i.e., years), the different angles of the segments (that is, the different multidimensionality of the test) may be important.

Other Considerations

Longitudinal Versus Cross-Sectional Cross-Grade Scales

At the Institute of Education Sciences (IES) Research Conference in June 2010, Jaekyung Lee of the University of Buffalo made an informal presentation of his current research comparing cross-grade scales from several of the K–12 testing companies and statewide assessments with the cross-grade scale from the Early Childhood Longitudinal Study-Kindergarten (ECLS-K). The ECLS-K cross-grade scales are unusual in that they are based on longitudinal item response data; other cross-grade scales, like NAEP's, have been constructed using cross-sectional data. Lee noted that the ECLS-K average growth curves are steeper than all of the others, which are generally similar to each other.

Once this is pointed out, it is easy to understand, at least superficially: Over the past couple decades, there has been a regular secular trend in test scores, with averages increasing over time. So if one compares scores (on the same test) in grades 4 and 5 in 2008, the difference is d_1, the cross-sectional difference. However, if one collects longitudinal data and compares scores of the same students on the same test in grades 4 and 5 (say in 2008 and then 2009), the difference is $d_1 + d_2$, where d_2 is the increase in educational performance between 2008 and 2009.

At first blush this may seem like nitpicking. However, a common use of cross-grade scales is to interpret them as a source of a value for "expected growth" between grades—and then individual (longitudinal) growth is compared to that "expected growth." If the "expected growth" is cross sectional, and longitudinal growth is what is being compared, it becomes complicated. It is complicated either way—the secular trend may not go on forever.

Curiously, the data analysis creating the longitudinal ECLS-K scales treated the longitudinal data as though they were cross sectional, and used essentially the same technology to analyze the data as has been used with NAEP cross-sectional data (Najarian, Pollack, Sorongon, & Hausken, 2009; Pollack, Atkins-Burnett, Najarian, & Rock, 2006; Pollack, Rock, Weiss, & Atkins-Burnett, 2005; Rock & Pollack, 2002). It may now be (just barely) possible to perform the computations that would be involved in the construction of a longitudinal cross-grade scale treating the repeated measurements as repeated (Cai, 2010). To do so would require longitudinal data collection.

Nevertheless, the standard of interest for policymakers is probably cross-sectional growth. The secular trend would probably be labeled "progress." This means that cross-sectional data collection, which is also quicker and easier, would answer the right question to provide meaning for points on the score scale. (If one wanted to fit individual curves to growth in achievement, however, longitudinal data would be required.)

Learning Progressions and Longitudinal Data Collection

"Learning progressions are logically and empirically derived sequences describing the way that knowledge and skill development typically occurs in a domain." There is increasing interest in the inclusion of items that measure progress through such learning progressions on achievement tests; for example, the plans of the Summative Multi-state Assessment Resources for Teachers and Educational Researchers (SMARTER) Balanced Assessment Consortium specifically mention the idea that some items may reflect learning progressions (SMARTER Balanced Assessment Consortium, n.d.).

The extent to which NAEP items might measure progress through learning progressions is not clear, because learning progressions have not been part of the frameworks for developing the assessment. However, if in the future learning progressions were to become part of the NAEP frameworks and assessment design, this would intersect with cross-grade scaling in two areas: interpretation and data collection.

With respect to interpretation, if responses to items, or a series of items, indicated progress through a sequence that had been established to be a learning progression, then that would establish a basis for across-grade score comparability: If items representing learning progressions made up a sufficiently large proportion of the assessment, scores could be interpreted to represent positions in those sequences, and could, hypothetically, be comparable for students in grades 4 and 8.

With respect to data collection, *some* longitudinal data collection would be required, at some point in the process, to establish that items (or more properly, item responses) represent or indicate progress on a sequence—that is, that the sequence has an empirical (as well as a logical) basis. This is not to say that the assessment itself must be longitudinal; cross-sectional data on an assessment made up of items that had been established to represent progress on some learning progression would yield information about the relative positions of the cross-sectional sample on that progression. However, at some point in the framework and item development processes, longitudinal data would be required to show that there is some degree of invariance in the putative order of the item responses; otherwise there could be counterclaims that the claimed "sequence" is not "sequential."

Conditional "One Year's Growth"

If "one year's growth" is to be useful as a value that makes points on the score scale more meaningful, it would be useful to know if the empirical average value of "one year's growth" is very different for students at different levels of the score scale or

from different demographic backgrounds. Because values for "one year's growth" are not currently available, the answers to these questions are unknown.

However, values for the difference between NAEP results at grades 4 and 8 are available, and tabulated in various ways in Appendix A (for the 2009 reading assessment, for which the cross-grade scale is endorsed), and Appendix B (for the 2009 mathematics assessment, with its historical, but not maintained, cross-grade scale). Those values for "four years' growth" give hints about what would happen if we knew "one year's growth."

For reading, the values of growth from grade 4 to grade 8 are remarkably consistent across the range from low-performing (10th percentile) to high-performing (90th percentile) students, and across most demographic groups: They are largely in the range 40–45 points. There is a slight (and perhaps surprising) tendency for lower-performing students to make larger gains.

The exceptions for reading are curious: Relatively high-performing students with disabilities show smaller gains (around 35 points) between grades 4 and 8, students in the Asian/Pacific Islander demographic classification exhibit slightly less growth (39 points for some levels), and Native American students show remarkably large gains from grades 4 to 8. (The latter may be associated with the very low scores for the lowest performing groups at grade 4.) English language learner (ELL) students show the smallest gains—around 30 points instead of the 40-something point gains that are typical for most.

The picture is different for mathematics. A regular feature across all of the tables in Appendix B is that high-scoring students exhibit larger gains between grades 4 and 8—for all students the difference is 54 points at the 90th percentile versus 34 points at the 10th percentile. Aside from this, the pattern across demographic groups is similar to that for reading, except that the Asian/Pacific Islander group does not have lower than average gains.

This all suggests that, for reading, one year's growth may be fairly uniform across students, with a few unsurprising exceptions. What the results might be for mathematics is less clear: It may be that mathematics "builds on itself," so that "the rich get richer" and large gains go with high scores. Or it may be that the unmaintained NAEP cross-grade mathematics scale is not the best way to examine the question.

How Fast Might "One Year's Growth" Change?

If a(n expensive) special study is done to estimate "one year's growth" on the NAEP scale, how often would it have to be repeated to remain accurate? This is much the same question that arises when other assessments are linked with NAEP, and the answer is probably the same: not every administration, but reasonably often.

Some clues can be obtained from NAEP results over time. For reading, there has been relatively little change in average scale score between 2002 and 2009, in either grade 4 (a 2-point increase, from 219 to 221) or grade 8 (no change, 264 both years).

The difference between the changes at the two grade levels means that, over seven years, "one year's growth" must have dropped slightly, but it cannot have dropped by more than a point per year. This suggests that the value of "one year's growth" for reading may be useful for a decade or more.

For mathematics, the clues are slightly less clear: There has been a relatively large change between 2000 and 2009 in grade 4 (a 14-point increase, from 226 to 240), and slightly less change at grade 8 (a 10-point increase, from 273 to 283). That is, the 2009 values for average scores at grades 4 and 8 are 4 points closer together than the 2000 values. The difference between the 2000 and 2009 changes for the two grade levels means that, over nine years, "one year's growth" must have dropped by a four-year aggregate value of 4 points. Given that the curve is presumably decelerating, the change in "one year's growth" across this time span likely exceeds 1 point for the lowest grade transitions (i.e., between 4th and 5th grade or between 5th and 6th grade). (In order to lose 4 points' gain in four years in the context of a decelerating curve, the decrease in one year's growth has to be somewhat more than 1 point per grade transition in the lower region of the curve and correspondingly somewhat less than 1 point per grade transition in the higher region of the curve.)

More detailed analysis of historical trends could make this estimate more precise, but it appears that once every 10 years may be a reasonable guess about how often the value of "one year's growth" should be checked.

Conclusion

Validity evidence can and should be assembled to support, and make more precise, interpretive statements of the first kind ("one year's growth"). "How many NAEP scale points is one year's growth?" is a question users of the scores can sensibly ask; there should be an answer. It is not difficult to obtain the answer; it is merely expensive. Samples of 3rd- and 5th-grade students could augment routine data collection for grade 4 to estimate growth on either side of grade 4. A sample of seventh-grade students could augment the grade 8 data collection. (Due to the common transition to high school at grade 9, it is not clear that cross-sectional data collection across the grade 8–9 boundary would be as informative, and defining a national probability sample across that boundary would be much more challenging.)

Interpretive statements of the second kind (cross-group comparisons across four-year spans) should probably be more actively discouraged. Even if useful cross-grade scales are maintained, there are sufficient grounds to argue that students in grades 4 and 8 with the same score exhibit different achievement. Such arguments reduce the usefulness of statements such as "this low-scoring group of 8th-grade students is similar to average 4th-grade students," and could be a distraction from more useful interpretations of the scores—in terms of achievement levels or item maps, for example.

These conclusions are consonant with the current official NAEP statement about the cross-grade scale for the 2009 reading assessment, which could be succinctly summarized as "it is cross grade, but don't push it."

References

Beaton, A. E. (Ed.) (1987). *Implementing the new design: The NAEP 1983–84 technical report* (No. 15-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Beaton, A. E. (Ed.) (1988). *Expanding the new design: The NAEP 1985–86 technical report* (No. 17-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). *Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions*. New York: MDRC, MDRC Working Papers on Research Methodology.

Briggs, D. (2010, July). *Why Aren't States with Developmental Score Scales Measuring Growth?* Presentation at the International Meeting of the Psychometric Society, Athens, GA.

Cai, L. (2010, July). *A two-tier full-information item factor analysis model with applications*. Presentation at the International Meeting of the Psychometric Society, Athens, GA.

Daro, P., & Shepard, L. (2011, May 20). *Memo to the NAEP Validity Studies Panel Re: Possible white paper on learning progressions and NAEP, and possible "pilot" study with sensitivity to instruction data set*. [Memorandum].

Haertel, E. H. (1991). *Report on TRP analyses of issues concerning within-age versus cross-age scales for the National Assessment of Educational Progress*. Washington, DC: National Center for Educational Statistics (available at http://www.eric.ed.gov:80/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED404367&ERICExtSearch_SearchType_0=no&accno=ED404367).

Johnson, E. G., & Allen, N. L. (Eds.) (1992). *The NAEP 1990 technical report* (Rep. No. 21-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Johnson, E. G., & Carlson, J. E. (Eds.) (1994). *The NAEP 1992 technical report* (Rep. No. 23-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Kolstad, A. (2004). *Issues in cross-grade scaling for the 2009 reading assessment*. Presentation at the Technical Panel Meeting to Discuss the Implementation of Within- and Cross-Grade Scaling for the NAEP 2009 Reading Assessment, Washington, DC, October 29.

McClellan, C. A., Donoghue, J. R., Gladkova, L., & Xu, X. (2005, November). *Cross-grade scales in NAEP: Research and real-life experience.* Presentation at the

conference Longitudinal Modeling of Student Achievement, Maryland Assessment Research Center for Education Success, University of Maryland, College Park, MD. Retrieved from http://www.education.umd.edu/EDMS/MARCES/conference/Longitudinal/Mcclellan.ppt

Najarian, M., Pollack, J. M., Sorongon, A. G., & Hausken, E. G. (2009). *Early Childhood Longitudinal Study, Kindergarten class of 1998–99 (ECLS-K), psychometric report for the eighth grade* (NCES 2009-002). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Patz, R. J. (2004). *Comments on item response theory in NAEP and vertical scaling.* Presentation at the Technical Panel Meeting to Discuss the Implementation of Within- and Cross-Grade Scaling for the NAEP 2009 Reading Assessment, Washington, DC, October 29.

Pollack, J. M., Atkins-Burnett, S., Najarian, M., and Rock, D. A. (2006). *Early Childhood Longitudinal Study, Kindergarten class of 1998–99 (ECLS-K), psychometric report for the fifth grade* (NCES 2006–036rev). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Pollack, J. M., Rock, D. A., Weiss, M., & Atkins-Burnett, S. (2005). *Early Childhood Longitudinal Study, Kindergarten CLASS of 1998–99 (ECLS-K), psychometric report for the third grade* (NCES 2005–062). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Reckase, M. D. (2004, October 29). *The vertical scaling of science achievement tests.* Presentation at the Technical Panel Meeting to Discuss the Implementation of Within- and Cross-Grade Scaling for the NAEP 2009 Reading Assessment, Washington, DC.

Rock, D. A., & Pollack, J. (2002). *Early Childhood Longitudinal Study, Kindergarten class of 1998–99 (ECLS-K), psychometric report for kindergarten through first grade* (NCES 2002–05). Washington, DC: National Center for Education Statistics, U.S. Department of Education.

SMARTER Balanced Assessment Consortium. (n.d.). *A summary of core components.* Retrieved from http://www.k12.wa.us/smarter/pubdocs/SBACSummary2010.pdf

Tirre, B. & Oranje, A. (2010, December 16). *NAEP Design Summit: Adjacent grades study*. Presentation at the NAEP Design Summit, Washington, DC.

Wise, L., & Hoffman, R. G. (2004). *Technical panel meeting to discuss the implementation of within- and cross-grade scaling for the NAEP 2009 Reading Assessment: Meeting notes* (DFR-04-74). Alexandria, VA: Human Resources Research Organization.

Yen, W. M. (2004, October 29). *Vertical (cross-grade) scaling*. Presentation at the Technical Panel Meeting to Discuss the Implementation of Within- and Cross-Grade Scaling for the NAEP 2009 Reading Assessment, Washington, DC.

Yen, W. M. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 233–251). New York: Springer.

Appendix A. NAEP: 2009 Reading Assessment
National Public + Private

Table A-1. Reading gap between 8th and 4th grade for all students

All students	10th percentile	25th percentile	50th percentile	75th percentile	90th percentile	Standard deviation
All students: 8th grade	219	243	267	288	305	34
Standard error	(0.5)	(0.4)	(0.3)	(0.4)	(0.4)	(0.2)
All students: 4th grade	175	199	223	245	264	35
Standard error	(0.5)	(0.4)	(0.3)	(0.3)	(0.3)	(0.2)
Gap: 8th grade – 4th grade	**44**	**44**	**44**	**43**	**41**	

Table A-2. Reading gap between 8th and 4th grade by disability classification

SD	10th percentile	25th percentile	50th percentile	75th percentile	90th percentile	Standard deviation
SD: 8th grade	178	202	229	253	274	38
Standard error	(1.7)	(1.2)	(0.6)	(0.9)	(0.5)	(0.5)
SD: 4th grade	132	159	189	217	241	42
Standard error	(1.1)	(1.0)	(0.7)	(0.7)	(0.7)	(0.4)
Gap: 8th grade – 4th grade	**45**	**43**	**39**	**36**	**34**	

Not SD	10th percentile	25th percentile	50th percentile	75th percentile	90th percentile	Standard deviation
Not SD: 8th grade	226	247	269	289	307	32
Standard error	(0.3)	(0.3)	(0.3)	(0.3)	(0.4)	(0.2)
Not SD: 4th grade	181	203	226	247	265	33
Standard error	(0.6)	(0.4)	(0.3)	(0.3)	(0.3)	(0.2)
Gap: 8th grade – 4th grade	**45**	**45**	**43**	**42**	**42**	

Table A-3. Reading gap between 8th and 4th grade by school lunch program eligibility

Eligible	10th percentile	25th percentile	50th percentile	75th percentile	90th percentile	Standard deviation
Eligible: 8th grade	205	229	251	272	289	34
Standard error	(0.8)	(0.4)	(0.3)	(0.5)	(0.4)	(0.2)
Eligible: 4th grade	161	185	208	230	248	34
Standard error	(0.7)	(0.5)	(0.3)	(0.3)	(0.3)	(0.2)
Gap: 8th grade – 4th grade	**43**	**44**	**43**	**42**	**42**	

Not eligible	10th percentile	25th percentile	50th percentile	75th percentile	90th percentile	Standard deviation
Not eligible: 8th grade	234	255	275	294	310	31
Standard error	(0.6)	(0.3)	(0.3)	(0.4)	(0.4)	(0.3)
Not eligible: 4th grade	192	213	234	254	271	32
Standard error	(0.5)	(0.3)	(0.3)	(0.4)	(0.6)	(0.2)
Gap: 8th grade – 4th grade	**42**	**41**	**41**	**40**	**40**	

Table A-4. Reading gap between 8th and 4th grade by race/ethnicity

White	**10th percentile**	**25th percentile**	**50th percentile**	**75th percentile**	**90th percentile**	**Standard deviation**
White: 8th grade	233	254	275	294	310	31
Standard error	(0.4)	(0.3)	(0.2)	(0.3)	(0.4)	(0.2)
White: 4th grade	190	211	232	252	269	32
Standard error	(0.3)	(0.4)	(0.3)	(0.3)	(0.4)	(0.2)
Gap: 8th grade – 4th grade	**44**	**43**	**42**	**42**	**41**	

Black	**10th percentile**	**25th percentile**	**50th percentile**	**75th percentile**	**90th percentile**	**Standard deviation**
Black: 8th grade	204	226	248	269	286	33
Standard error	(0.7)	(0.6)	(0.5)	(0.6)	(0.5)	(0.2)
Black: 4th grade	161	184	206	228	246	33
Standard error	(0.8)	(0.7)	(0.6)	(0.5)	(0.7)	(0.3)
Gap: 8th grade – 4th grade	**43**	**43**	**42**	**41**	**40**	

Hispanic	**10th percentile**	**25th percentile**	**50th percentile**	**75th percentile**	**90th percentile**	**Standard deviation**
Hispanic: 8th grade	203	228	252	273	291	35
Standard error	(1.9)	(0.7)	(0.7)	(0.6)	(0.5)	(0.5)
Hispanic: 4th grade	159	183	208	229	248	35
Standard error	(0.9)	(0.7)	(0.7)	(0.7)	(1.0)	(0.4)
Gap: 8th grade – 4th grade	**44**	**45**	**45**	**44**	**42**	

Asian/Pacific Island	**10th percentile**	**25th percentile**	**50th percentile**	**75th percentile**	**90th percentile**	**Standard deviation**
Asian/Pacific Island: 8th grade	229	254	277	298	316	35
Standard error	(2.1)	(1.1)	(1.1)	(1.3)	(1.8)	(0.6)
Asian/Pacific Island: 4th grade	190	214	237	259	277	35
Standard error	(2.2)	(1.6)	(1.2)	(1.7)	(1.4)	(0.6)
Gap: 8th grade – 4th grade	**39**	**41**	**40**	**39**	**39**	

American Indian	**10th percentile**	**25th percentile**	**50th percentile**	**75th percentile**	**90th percentile**	**Standard deviation**
American Indian: 8th grade	205	229	253	276	296	37
Standard error	(1.8)	(2.1)	(1.4)	(2.0)	(1.6)	(0.7)
American Indian: 4th grade	148	178	208	232	253	41
Standard error	(3.0)	(2.9)	(1.4)	(1.7)	(2.2)	(1.0)
Gap: 8th grade – 4th grade	**56**	**51**	**45**	**44**	**43**	

Table A-5. Reading gap between 8th and 4th grade by gender

Male	10th percentile	25th percentile	50th percentile	75th percentile	90th percentile	Standard deviation
Male: 8th grade	214	239	262	283	301	35
Standard error	(0.6)	(0.3)	(0.3)	(0.3)	(0.4)	(0.2)
Male: 4th grade	170	196	220	243	261	36
Standard error	(0.6)	(0.4)	(0.3)	(0.3)	(0.7)	(0.3)
Gap: 8th grade – 4th grade	**44**	**43**	**42**	**41**	**40**	

Female	10th percentile	25th percentile	50th percentile	75th percentile	90th percentile	Standard deviation
Female: 8th grade	225	248	271	292	309	33
Standard error	(0.5)	(0.4)	(0.4)	(0.5)	(0.5)	(0.2)
Female: 4th grade	180	203	226	248	266	34
Standard error	(0.7)	(0.5)	(0.3)	(0.3)	(0.4)	(0.2)
Gap: 8th grade – 4th grade	**46**	**45**	**44**	**44**	**43**	

Table A-6. Reading gap between 8th and 4th grade by ELL status

ELL	10th percentile	25th percentile	50th percentile	75th percentile	90th percentile	Standard deviation
ELL: 8th grade	173	197	221	244	263	36
Standard error	(1.0)	(2.5)	(1.4)	(1.5)	(2.2)	(0.7)
ELL: 4th grade	142	166	190	212	230	35
Standard error	(1.2)	(0.9)	(1.2)	(1.0)	(0.9)	(0.6)
Gap: 8th grade – 4th grade	**31**	**31**	**31**	**32**	**33**	

Not ELL	10th percentile	25th percentile	50th percentile	75th percentile	90th percentile	Standard deviation
Not ELL: 8th grade	224	246	268	289	306	33
Standard error	(0.4)	(0.3)	(0.3)	(0.3)	(0.4)	(0.1)
Not ELL: 4th grade	180	203	226	247	265	34
Standard error	(0.4)	(0.3)	(0.3)	(0.3)	(0.3)	(0.2)
Gap: 8th grade – 4th grade	**44**	**43**	**42**	**41**	**41**	

Appendix B. NAEP: 2009 Mathematics Assessment National Public + Private

Table B-1. Mathematics gap between 8th and 4th grade for all students

All students	10th percentile	25th percentile	50th percentile	75th percentile	90th percentile	Standard deviation
All students: 8th grade	236	259	284	308	329	36
Standard error	(0.50)	(0.30)	(0.30)	(0.40)	(0.50)	(0.20)
All students: 4th grade	202	221	241	260	275	29
Standard error	(0.40)	(0.30)	(0.30)	(0.30)	(0.20)	(0.10)
Gap: 8th grade – 4th grade	**34**	**38**	**43**	**48**	**54**	

Table B-2. Mathematics gap between 8th and 4th grade by disability classification

SD	10th percentile	25th percentile	50th percentile	75th percentile	90th percentile	Standard deviation
SD: 8th grade	199	221	246	270	293	37
Standard error	(1.4)	(1.1)	(0.7)	(0.8)	(0.8)	(0.4)
SD: 4th grade	178	198	221	242	260	32
Standard error	(0.8)	(0.9)	(0.5)	(0.6)	(0.8)	(0.3)
Gap: 8th grade – 4th grade	**21**	**23**	**26**	**29**	**34**	

Not SD	10th percentile	25th percentile	50th percentile	75th percentile	90th percentile	Standard deviation
Not SD: 8th grade	242	264	287	310	330	34
Standard error	(0.3)	(0.4)	(0.4)	(0.4)	(0.4)	(0.2)
Not SD: 4th grade	206	224	243	261	276	27
Standard error	(0.4)	(0.3)	(0.3)	(0.3)	(0.3)	(0.1)
Gap: 8th grade – 4th grade	**36**	**40**	**44**	**49**	**54**	

Table B-3. Mathematics gap between 8th and 4th grade by school lunch program eligibility

Eligible	10th percentile	25th percentile	50th percentile	75th percentile	90th percentile	Standard deviation
Eligible: 8th grade	222	244	268	290	309	34
Standard error	(0.8)	(0.4)	(0.4)	(0.4)	(0.4)	(0.2)
Eligible: 4th grade	192	210	229	246	261	27
Standard error	(0.6)	(0.3)	(0.3)	(0.2)	(0.3)	(0.2)
Gap: 8th grade – 4th grade	**31**	**34**	**39**	**43**	**48**	

Not eligible	10th percentile	25th percentile	50th percentile	75th percentile	90th percentile	Standard deviation
Not eligible: 8th grade	250	272	295	317	336	34
Standard error	(0.5)	(0.4)	(0.3)	(0.3)	(0.4)	(0.2)
Not eligible: 4th grade	216	234	251	268	282	26
Standard error	(0.5)	(0.3)	(0.3)	(0.3)	(0.4)	(0.2)
Gap: 8th grade – 4th grade	**34**	**38**	**43**	**49**	**54**	

Table B-4. Mathematics gap between 8th and 4th grade by race/ethnicity

White	**10th percentile**	**25th percentile**	**50th percentile**	**75th percentile**	**90th percentile**	**Standard deviation**
White: 8th grade	251	272	294	315	334	33
Standard error	(0.5)	(0.4)	(0.4)	(0.3)	(0.4)	(0.2)
White: 4th grade	215	232	249	266	280	26
Standard error	(0.4)	(0.2)	(0.2)	(0.3)	(0.3)	(0.1)
Gap: 8th grade – 4th grade	**36**	**40**	**44**	**50**	**55**	

Black	**10th percentile**	**25th percentile**	**50th percentile**	**75th percentile**	**90th percentile**	**Standard deviation**
Black: 8th grade	218	239	262	283	303	33
Standard error	(0.8)	(0.7)	(0.4)	(0.5)	(0.8)	(0.3)
Black: 4th grade	187	205	223	241	256	27
Standard error	(0.4)	(0.5)	(0.5)	(0.4)	(0.4)	(0.2)
Gap: 8th grade – 4th grade	**31**	**34**	**38**	**43**	**47**	

Hispanic	**10th percentile**	**25th percentile**	**50th percentile**	**75th percentile**	**90th percentile**	**Standard deviation**
Hispanic: 8th grade	222	244	268	290	310	34
Standard error	(1.2)	(0.9)	(0.7)	(0.5)	(0.7)	(0.4)
Hispanic: 4th grade	192	210	229	246	261	27
Standard error	(1.0)	(0.5)	(0.6)	(0.3)	(0.6)	(0.4)
Gap: 8th grade – 4th grade	**30**	**34**	**39**	**44**	**49**	

Asian /Pacific Island	**10th percentile**	**25th percentile**	**50th percentile**	**75th percentile**	**90th percentile**	**Standard deviation**
Asian/Pacific Island: 8th grade	252	277	303	327	347	37
Standard error	(1.5)	(2.1)	(1.9)	(1.4)	(2.2)	(0.6)
Asian/Pacific Island: 4th grade	217	237	256	274	291	29
Standard error	(1.0)	(1.5)	(1.1)	(1.0)	(1.5)	(0.5)
Gap: 8th grade – 4th grade	**36**	**40**	**46**	**52**	**57**	

American Indian	**10th percentile**	**25th percentile**	**50th percentile**	**75th percentile**	**90th percentile**	**Standard deviation**
American Indian: 8th grade	217	241	268	291	313	38
Standard error	(2.3)	(2.9)	(0.8)	(0.9)	(1.2)	(0.8)
American Indian: 4th grade	186	206	226	246	262	29
Standard error	(2.0)	(2.4)	(1.3)	(0.8)	(1.7)	(0.5)
Gap: 8th grade – 4th grade	**31**	**35**	**41**	**45**	**51**	

Table B-5. Mathematics gap between 8th and 4th grade by gender

Male	10th percentile	25th percentile	50th percentile	75th percentile	90th percentile	Standard deviation
Male: 8th grade	236	260	285	310	331	37
Standard error	(0.6)	(0.4)	(0.4)	(0.4)	(0.5)	(0.2)
Male: 4th grade	202	222	242	261	277	30
Standard error	(0.6)	(0.4)	(0.3)	(0.3)	(0.4)	(0.2)
Gap: 8th grade – 4th grade	**34**	**38**	**43**	**49**	**54**	

Female	10th percentile	25th percentile	50th percentile	75th percentile	90th percentile	Standard deviation
Female: 8th grade	236	259	283	306	327	35
Standard error	(0.6)	(0.4)	(0.4)	(0.4)	(0.5)	(0.2)
Female: 4th grade	202	221	240	258	273	28
Standard error	(0.4)	(0.3)	(0.3)	(0.3)	(0.3)	(0.2)
Gap: 8th grade – 4th grade	**34**	**38**	**43**	**48**	**53**	

Table B-6. Mathematics gap between 8th and 4th grade by ELL status

ELL	10th percentile	25th percentile	50th percentile	75th percentile	90th percentile	Standard deviation
ELL: 8th grade	200	221	243	265	285	34
Standard error	(1.9)	(2.2)	(1.2)	(1.8)	(1.5)	(0.7)
ELL: 4th grade	182	200	219	237	252	28
Standard error	(1.3)	(1.2)	(0.7)	(0.9)	(0.7)	(0.6)
Gap: 8th grade – 4th grade	**18**	**21**	**25**	**29**	**33**	

Not ELL	10th percentile	25th percentile	50th percentile	75th percentile	90th percentile	Standard deviation
Not ELL: 8th grade	240	262	286	309	330	35
Standard error	(0.4)	(0.3)	(0.3)	(0.4)	(0.4)	(0.2)
Not ELL: 4th grade	206	224	243	261	276	28
Standard error	(0.3)	(0.3)	(0.2)	(0.3)	(0.3)	(0.1)
Gap: 8th grade – 4th grade	**34**	**38**	**43**	**48**	**53**	